POLITECHNIKA GDAŃSKA
Wydział Elektroniki, Telekomunikacji i Informatyki

Michał Lech

Metoda i algorytmy sterowania procesami miksowania dźwięku za pomocą gestów w oparciu o analizę obrazu wizyjnego

Rozprawa doktorska

Promotor:
prof. dr hab. inż. Bożena Kostek, prof. zw. PG
Wydział Elektroniki, Telekomunikacji i Informatyki
Politechnika Gdańska

Gdańsk, 2012

Podziękowania

Autor rozprawy pragnie wyrazić słowa podziękowania Pani Promotor, prof. dr hab. inż. Bożenie Kostek za wszelkie uwagi udzielone w trakcie prac nad rozprawą, Kierownikowi Katedry Systemów Multimedialnych, prof. dr. hab. inż. Andrzejowi Czyżewskiemu za stworzenie i zapewnienie możliwości rozwoju zainteresowań naukowych oraz zaproszonym inżynierom dźwięku i Kolegom z Katedry za udział w testach subiektywnych.

Algorytmy zastosowane w systemie przedstawionym w rozprawie zostały częściowo opracowane w ramach projektu: POIG.01.03.01-22-017/08 pt.: "Opracowanie typoszeregu komputerowych interfejsów multimodalnych oraz ich wdrożenie w zastosowaniach edukacyjnych, medycznych, w obronności i w przemyśle". Projekt jest współfinansowany ze środków Europejskiego Funduszu Rozwoju Regionalnego i budżetu państwa.

Słownik pojęć

znaczenie wybranych terminów i skrótów

(w porządku alfabetycznym)

Termin	Opis / wyjaśnienie
balans	różnica pomiędzy poziomami poszczególnych sygnałów fonicznych tworzących miks; w zbalansowanym nagraniu zachowane są równomierne proporcje głośności poszczególnych dźwięków
BGR (ang. *Blue*, *Green*, *Red*)	reprezentacja modelu przestrzeni barw RGB, w którym składowa niebieska jest zamieniona miejscami ze składową czerwoną
BGRA (ang. *Blue*, *Green*, *Red*, *Alpha*)	reprezentacja modelu przestrzeni barw RGBA, w którym poza komponentami RGB występuje dodatkowo kanał alfa umożliwiający uzyskanie przezroczystości
CSS (ang. *Curvature Scale Space*)	metoda przestrzeni krzywizn skalowalnych
dynamika (w nagraniu muzycznym)	różnica pomiędzy dźwiękiem najcichszym i najgłośniejszym
FCL (ang. *Fuzzy Control Language*)	dziedzinowy język opisu systemu wnioskowania rozmytego zgodny ze standardem IEC 61131-7
głębia (w nagraniu muzycznym)	parametr charakteryzujący subiektywne wrażenie odległości słuchacza od poszczególnych dźwięków (instrumentów) w miksie; głębia kreowana jest przez wykorzystanie pogłosu (naturalnego bądź sztucznego)
grid-search	metoda przeszukiwania zbioru parametrów algorytmu z dziedziny uczenia maszynowego w celu znalezienia parametrów optymalnych
HCI (ang. *Human-Computer Interaction*)	wzajemne oddziaływanie pomiędzy człowiekiem a komputerem, zachodzące poprzez interfejs użytkownika
JNI (ang. *Java Native Interface*)	interfejs programistyczny dla języka Java umożliwiający wywoływanie w jego obrębie funkcji kodu C, C++ i asembler
kompresor dynamiki	urządzenie (lub algorytm) realizujące zmniejszenie dynamiki sygnału (tj. zmniejszenie różnicy poziomów dźwięku najgłośniejszego i najcichszego) w stosunku do dynamiki sygnału wejściowego
korektor częstotliwości	zespół filtrów, które służą do wzmacniania lub tłumienia określonego zakresu częstotliwości, zmieniając tym samym barwę dźwięku
LIBSVM	biblioteka programistyczna udostępniająca imple-

	mentację maszyny wektorów nośnych
MIDI (ang. *Musical Instrument Digital Interface*)	interfejs wymiany informacji pomiędzy elektronicznymi instrumentami i urządzeniami muzycznymi
miksowanie dźwięku	*telekom.* składanie (mieszanie) sygnałów fonicznych (pochodzących z kilku niezależnych źródeł) prowadzące do powstania jednego sygnału wypadkowego; wiąże się zwykle ze zmianą parametrów sygnałów składowych (np. poziomu głośności, widma częst.); przeprowadzane w mikserze (dźwięku); stosowane w technice radiofonicznej, film., fonograficznej [118]
OpenCV	biblioteka udostępniająca funkcje przetwarzania danych (w szczególności obrazu), opracowana przez firmę Intel
oprogramowanie DAW (ang. *Digital Audio Workstation*)	oprogramowanie cyfrowej stacji roboczej umożliwiającej nagrywanie, edycję i miksowanie dźwięku
panorama	przestrzeń stereofoniczna w nagraniu fonicznym
plug-in (wtyczka)	oprogramowanie działające w środowisku systemu DAW realizujące przetwarzanie dźwięku
pogłos	zjawisko stopniowego zanikania energii dźwięku po ucichnięciu źródła, związane z występowaniem dużej liczby fal odbitych od powierzchni pomieszczenia
poziom (w nagraniu muzycznym)	poziom natężenia dźwięku; logarytmiczna miara natężenia dźwięku w stosunku do pewnej umownie przyjętej wartości odniesienia, wyrażana w decybelach
punkt optymalnego odsłuchu (ang. *sweet-spot*)	miejsce w przestrzeni, w którym dźwięk jest słyszany zgodnie z intencją inżyniera dźwięku lub projektanta systemu odsłuchowego
realizator nagrań	osoba odpowiedzialna w studiu nagrań za przebieg czynności wchodzących w skład etapu produkcji nagrania fonicznego; z reguły, odpowiedzialna również za przebieg procesów etapu postprodukcji, a w szczególności miksowania; w pełni profesjonalnej produkcji (i postprodukcji) muzycznej procesy miksowania powierza się inżynierowi miksowania, natomiast realizator nagrań dba jedynie o prawidłowy przebieg nagrania; w niniejszej rozprawie pod pojęciem realizatora nagrań rozumie się również osobę miksującą nagranie
RGB (ang. *Red, Green, Blue*)	jeden z modeli przestrzeni barw; model wynikający z właściwości odbiorczych ludzkiego oka, w którym wrażenie widzenia dowolnej barwy można wywołać przez zmieszanie w ustalonych proporcjach trzech wiązek światła o barwie czerwonej, zielonej i niebie-

	skiej
SVM (ang. *Support Vector Machine*)	maszyna wektorów nośnych; klasyfikator bazujący na hiperpłaszczyźnie rozdzielającej z maksymalnym marginesem obiekty należące do dwóch klas
Swing	biblioteka platformy Java SE udostępniająca elementy tworzenia interfejsu graficznego
szerokość bazy	parametr, który określa rozmiar sceny, na której rozmieszczone są pozorne źródła dźwięku, z których dociera dźwięk do słuchacza; przy zbyt wąskiej bazie nagranie może sprawiać wrażenie monofonicznego; w przypadku zbyt szerokiej bazy można zauważyć podział sceny na lewą i prawą stronę
ścieżka	żargonowe, zwyczajowe określenie kontenera sygnałów fonicznych w oprogramowaniu DAW, z którym związany jest dany zbiór efektów modyfikujących dźwięk
ToF (ang. *Time-of-Flight*)	technika pomiaru czasu przemieszczenia się obiektu, cząstki lub fali w danym ośrodku; zastosowana w kamerach ToF umożliwia pozyskanie informacji o głębi w oparciu o pomiar czasu, jaki upływa od wysłania z kamery wiązki światła i powrotu do niej po odbiciu od obiektu
VRML (ang. *Virtual Reality Modelling Language*)	język modelowania wirtualnej rzeczywistości; standard formatu pliku opisującego grafikę 3D i interaktywną grafikę wektorową
WEKA	otwarte środowisko udostępniające algorytmy uczenia maszynowego oraz narzędzia do ich testowania
„zakolorowania" dźwięku	z reguły niepożądane zmiany w oryginalnej charakterystyce częstotliwościowej dźwięku objawiające się osłabieniem lub wzmocnieniem określonego fragmentu pasma; najczęściej wiązane ze zjawiskiem filtracji grzebieniowej
zgranie (miks)	suma sygnałów fonicznych

Wykaz oznaczeń

(wykaz najważniejszych oznaczeń w kolejności wystąpienia w rozprawie)

p_{ij}	piksel wchodzący w skład obrazu o wymiarach *i* x *j*, pozyskanego z kamery
p^{red}	piksel wchodzący w skład czerwonego obrazu kalibracyjnego
p^{green}	piksel wchodzący w skład zielonego obrazu kalibracyjnego
p^{blue}	piksel wchodzący w skład niebieskiego obrazu kalibracyjnego
p^{white}	piksel wchodzący w skład białego obrazu kalibracyjnego
p^{black}	piksel wchodzący w skład czarnego obrazu kalibracyjnego
p^{r}_{ij}	piksel wchodzący w skład profilu korekcji otrzymanego dla czerwonego obrazu kalibracyjnego
p^{g}_{ij}	piksel wchodzący w skład profilu korekcji otrzymanego dla zielonego obrazu kalibracyjnego
p^{b}_{ij}	piksel wchodzący w skład profilu korekcji otrzymanego dla niebieskiego obrazu kalibracyjnego
p^{wh}_{ij}	piksel wchodzący w skład profilu korekcji otrzymanego dla białego obrazu kalibracyjnego
p^{bk}_{ij}	piksel wchodzący w skład profilu korekcji otrzymanego dla czarnego obrazu kalibracyjnego
r_{ij}	czerwona składowa piksela p_{ij}
g_{ij}	zielona składowa piksela p_{ij}
b_{ij}	niebieska składowa piksela p_{ij}
p'_{ij}	piksel wchodzący w skład obrazu z kamery po zastosowaniu profilu korekcji barwowej
p^{c}_{ij}	piksel wchodzący w skład profilu korekcji barwowej
t^{rgb}	próg rozróżniania zabarwienia obrazu pomiędzy komponentami RGB
t^{wh}	próg rozróżniania białego obrazu
t^{bk}	próg rozróżniania czarnego obrazu
$\mathbf{p'}$	obraz z kamery po zastosowaniu profilu korekcji barwowej
$\mathbf{p}^{\mathbf{screen}}$	przetworzony obraz wyświetlany przez projektor

$\mathbf{p}_{out}$	obraz będący wynikiem odejmowania obrazu **p'** od obrazu $\mathbf{p}^{screen}$
p_{ij}^{gray}	piksel wchodzący w skład obrazu $\mathbf{p}_{out}$ poddanego operacji konwersji do percepcyjnie ważonej skali szarości
r_{ij}^{gray}	czerwona składowa obrazu utworzonego przez piksele p_{ij}^{gray}
g_{ij}^{gray}	zielona składowa obrazu utworzonego przez piksele p_{ij}^{gray}
b_{ij}^{gray}	niebieska składowa obrazu utworzonego przez piksele p_{ij}^{gray}
p_{ij}^{bin}	obraz złożony z pikseli p_{ij}^{gray} po binaryzacji
u_{ij}	wektor ruchu utworzony z pozycji ręki w obrazie w chwilach t_i i t_j
υ_{ij}	prędkość ruchu ręki dla wektora u_{ij}
x_i	współrzędna x pozycji ręki w obrazie w chwili t_i
x_j	współrzędna x pozycji ręki w obrazie w chwili t_j
y_i	współrzędna y pozycji ręki w obrazie w chwili t_i
y_j	współrzędna y pozycji ręki w obrazie w chwili t_j
α_{ij}	kąt pomiędzy wektorem u_{ij} a wersorem osi y
φ_{ij}	kąt związany relacją z kątem α_{ij}
υ_t^x	prędkość pozioma ruchu ręki w chwili t
υ_t^y	prędkość pionowa ruchu ręki w chwili t
f_{FR}	rozdzielczość czasowa systemu (prędkość przetwarzania w klatkach na sekundę)
$n_{\tau_0}^{\tau_1}$	interwał pomiędzy kluczowymi ramkami
p	prawdopodobieństwo testowe w testach statystycznych
SS Efekt	parametr wyznaczany w ramach testu Friedmana; suma kwadratów odchyleń pomiędzy grupami;
df Efekt	parametr wyznaczany w ramach testu Friedmana; liczba stopni swobody pomiędzy grupami
MS Efekt	parametr wyznaczany w ramach testu Friedmana; efekt średniokwadratowy, zmienność pomiędzy grupami
SS Błąd	parametr wyznaczany w ramach testu Friedmana; suma kwadratów odchyleń wewnątrz grup, prawdziwy błąd losowy;
df Błąd	parametr wyznaczany w ramach testu Friedmana; liczba stopni swobody wewnątrz grup

MS Błąd	parametr wyznaczany w ramach testu Friedmana; błąd średniokwadratowy, zmienność wewnątrz grup
χ^2	wartość testu zgodności chi-kwadrat, związana z rozkładem χ^2

Spis treści

1 Wprowadzenie 9
2 Percepcja i interakcja wielomodalna 15
3 Wybrane metody rozpoznawania gestów rąk 18
3.1 Wybrane metody rozpoznawania gestów statycznych 20
3.1.1 Metody ekstrakcji obszaru dłoni z obrazu 22
3.1.2 Metody parametryzacji dłoni 23
3.1.3 Metody rozpoznawania statycznych gestów dłoni oparte na modelach 26
3.1.4 Metody klasyfikacji gestów dłoni 28
3.2 Wybrane metody rozpoznawania gestów dynamicznych 29
3.2.1 Modelowanie gestów za pomocą stanów 30
3.2.2 Modelowanie ruchu 32
4 Wybrane metody i systemy kształtowania dźwięku za pomocą gestów 35
5 System miksowania dźwięku za pomocą gestów rąk 42
5.1 Założenia projektowe 42
5.1.1 Wymagania stawiane systemowi 43
5.2 Wybór metody klasyfikacji gestów rąk 45
5.2.1 Badanie skuteczności rozpoznawania gestów statycznych 46
5.2.2 Wyniki eksperymentów wstępnych 48
5.2.3 Wybór metod 53
5.3 Komponenty i architektura systemu 54
5.4 Interfejs graficzny 57
5.5 Słownik gestów 61
5.6 Implementacja systemu 64
5.7 Uruchomienie i kalibracja systemu 65
5.8 Zastosowane metody i algorytmy 69
5.8.1 Przetwarzanie wstępne obrazu 69
5.8.2 Metoda detekcji i śledzenia rąk 71
5.8.3 Metoda rozpoznawania gestów dynamicznych 77
5.8.4 Metoda rozpoznawania gestów statycznych 82
6 Badanie opracowanego systemu 85
6.1 Wpływ wizualizacji parametrów i ergonomii na wyniki procesu miksowania 85

6.1.1 Metodyka miksowania sygnałów 86
6.1.2 Warunki eksperymentu 88
6.1.3 Metodyka testów subiektywnych 88
6.1.4 Analiza wyników badań z udziałem realizatorów nagrań 92
6.1.5 Analiza wyników testów subiektywnych 108
6.2 Badanie wydajności systemu 114
6.3 Badanie skuteczności rozpoznawania gestów dynamicznych 116
7 Podsumowanie i wnioski 120
Bibliografia 126
Dodatek A. Wzór ankiety do oceny słownika gestów pod względem intuicyjności 140
Dodatek B. Definicja systemu rozmytego w kodzie FCL 146
Dodatek C. Wizualizacje zgrań i wartości parametrów 151
Dodatek D. Analiza statystyczna rozkładu wartości parametrów 172
Dodatek E. Analiza statystyczna przydzielonych przez ekspertów ocen walorów estetycznych zgrań 175
Dodatek F. Analiza statystyczna wartości skuteczności rozpoznawania gestów dynamicznych 177
Dodatek G. Dokumentacja techniczna systemu 179
Dodatek H. Ankieta dla realizatorów 206
Dodatek I. Dodatek DVD 209

1 Wprowadzenie

Miksowanie sygnałów fonicznych jest nieodłącznym elementem profesjonalnej produkcji muzycznej. Śledząc zmiany w sposobie miksowania dźwięku w nagraniu muzycznym na przestrzeni lat, zauważyć można, że w głównej mierze podyktowane one były rozwojem techniki. Wraz z rozwojem elektrycznego zapisu dźwięku i pojawieniem się rejestratorów wielościeżkowych umożliwiających odczyt synchronizowany z zapisem (*Selsync – Selective synchronous recording*) etap miksowania zaczął być powoli oddzielany od etapu nagrywania [109] [121]. Możliwe stało się tworzenie wielu wersji tego samego nagrania różniących się od siebie parametrami wpływającymi na subiektywny odbiór muzyki. Z kolei, możliwość odtwarzania nagrania z rejestratora o większej liczbie ścieżek i miksowania go z wykorzystaniem wielu kanałów stołu mikserskiego wykształciła technologię, w której parametry stołu mikserskiego modyfikowane były przez kilka osób jednocześnie [63] [109]. Konieczne było nauczenie się sekwencji wykonywanych czynności, a niewielki błąd mógł powodować potrzebę ich powtórzenia. Pojawienie się techniki cyfrowej i stołów mikserskich z automatyką umożliwiło miksowanie złożonych sesji produkcyjnych przez tylko jedną osobę [121]. Możliwe stało się również miksowanie nieliniowe [109]. Dalszy rozwój urządzeń fonicznych i komputeryzacja przyczyniły się do zwiększenia efektywności ugruntowanych technik miksowania i wykształcenia nowych sposobów działania. Pojawienie się rejestratorów ADAT (ang. *Alesis Digital Audio Tape*) umożliwiło muzykom realizację swoich nagrań w sposób półprofesjonalny w domu (ang. *homerecording*) przy niewielkim nakładzie finansowym [136].

W ostatnim dziesięcioleciu, dzięki znacznemu wzrostowi stosunku jakości sprzętu fonicznego do ceny, pojawiła się tendencja do wypierania dużych, bogato wyposażonych studiów nagrań przez studia niskobudżetowe. W studiach tych dominuje podejście polegające na miksowaniu nagrań w oparciu o oprogramowanie komputerowe DAW (ang. *Digital Audio Workstation*) bez wykorzystania stołu mikserskiego. Oczywiste uzasadnienie takiego stanu rzeczy stanowią względy ekonomiczne. Jednocześnie wielu znanych inżynierów dźwięku podkreśla fakt, że efekty miksowania nagrań z wykorzystaniem jedynie komputera (ang. *mixing in the box*) często nie są tak dobrze oceniane, jak w przypadku tradycyjnego podejścia, w którym wykorzystuje się stół mikserski [22] [32]

[90] [109]. Jako jeden z powodów podaje się między innymi różnice między algorytmami przetwarzania sygnału zaimplementowanymi w oprogramowaniu studyjnym a ich fizycznymi odpowiednikami w drogich, analogowych stołach mikserskich [22] [32] [43] [90] [91]. Według słów inżynierów dźwięku zajmujących się zawodowo pracą studyjną, „*plug-iny* mniej ubarwiają dźwięk, można powiedzieć, że są zbyt dokładne" [43], natomiast urządzenia analogowe „mają swoje brzmienie" [115]. Spotkać się można także z opiniami głoszącymi, że jakość algorytmów nie przekłada się znacząco na osiągane wyniki, natomiast finalne brzmienie wynika przede wszystkim z ergonomii interfejsu wykorzystywanego do miksowania [65] [66] [115]. Znaczenie kultury pracy i ergonomii podkreślają również inżynierowie dźwięku ceniący wyżej urządzenia analogowe od ich cyfrowych emulacji lub sprzętu cyfrowego [43] [65] [66]. Uniwersalność myszy i klawiatury, wykorzystywanych powszechnie do obsługi większości aplikacji komputerowych, spowodowała zaadaptowanie tego interfejsu również na potrzeby obsługi oprogramowania DAW. Poważnym ograniczeniem takiego interfejsu jest z reguły brak możliwości jednoczesnej edycji więcej niż jednego parametru. Modyfikowanie wartości parametrów za pomocą myszy nie jest również tak wygodne, jak w przypadku tradycyjnych interfejsów wyposażonych w regulatory potencjometryczne. Ograniczenia te były przyczyną opracowania szeregu nowych interfejsów dedykowanych do pracy z dźwiękiem. Cechami charakterystycznymi urządzeń tego typu jest powielanie elementów stołów mikserskich takich, jak potencjometry obrotowe i suwakowe lub mierniki diodowe przy jednoczesnym zachowaniu niewielkich gabarytów. Ograniczenie wielkości interfejsu realizowane jest przez możliwość przypisania danego potencjometru do wybranej funkcji oprogramowania. Zmiana powiązania potencjometru z funkcją może jednak wymagać użycia myszy [131]. Pomimo zwiększenia ergonomii pracy z wykorzystaniem takich interfejsów, ich cena stoi w sprzeczności z ideą nagrywania w studiach domowych. Przykładem rozwiązania tego problemu może być zaadaptowanie na potrzeby miksowania uniwersalnych interfejsów z innych dziedzin, zwiększających jednak ergonomię w porównaniu do myszy i klawiatury. Przykładem może być ekran dotykowy. Znane jest wykorzystanie produktu iPad [2] jako bezprzewodowego sterownika oprogramowania DAW [123] lub cyfrowych stołów mikserskich [116]. Przesunięcie wirtualnego suwaka na ekranie tabletu powoduje ruch odpowiadającego suwaka stołu mikserskiego lub miksera wirtualnego aplikacji DAW. Jednocześnie, producenci oprogra-

mowania DAW podejmują próby wykonania zabiegu odwrotnego, polegającego na dopasowaniu aplikacji do specyfiki interfejsu. Przykładem może być aplikacja Sonoma Wire Works StudioTrack pracująca bezpośrednio w środowisku urządzenia iPad.

Zastosowanie interfejsu dotykowego sprzyja dodatkowo opracowywaniu nowego rodzaju interfejsów graficznych mikserów wirtualnych. W systemie przedstawionym w filmie "*The art of mixing*" [42] źródła sygnałów fonicznych prezentowane są w postaci kulistych kształtów osadzonych w przestrzeni trójwymiarowej. Modyfikacja parametrów fonicznych odbywa sie poprzez zmianę pozycji, wielkości lub rozciągnięcia kształtu. Prezentowanie informacji w ten sposób ma również funkcję dydaktyczną, co wynika z możliwości łatwej, wzrokowej oceny zależności dźwiękowych pomiędzy źródłami. Z drugiej strony, inżynierowie dźwięku wskazują na negatywne skutki miksowania z udostępnioną jednocześnie informacją wizualną [36]. Przesłanie tego typu przekazał Steve Lilliwhite, światowej sławy inżynier dźwięku i producent muzyczny, w trakcie swojego wykładu plenarnego na 133 Konwencji Audio Engineering Society zatytułowanego: „***Listen with Your Ears and Not Your Eyes***". W przypadku widocznej informacji wizualnej inżynierowie dźwięku, w trakcie modyfikacji sygnałów fonicznych, mogą przywiązywać zbyt dużą wagę do informacji wyświetlanych na ekranie [36] [115]. Może to prowadzić do wykonywania zgrań o słabszych walorach estetycznych w porównaniu ze zgraniami tworzonymi w przypadku, kiedy decyzja jest podejmowana jedynie na podstawie oceny słuchowej. Jako główny powód podaje się fizjologię systemów sensorycznych i mechanizmy wielomodalnej percepcji, w której nadrzędną rolę pełni zmysł wzroku [4] [13]. Reprezentowanie zmian parametrów sygnałów fonicznych w postaci informacji wizualnej może również wpływać na percepcję dźwięku na niższych poziomach systemu sensorycznego. Przykładowo, znane są prace, w których potwierdzono istnienie wpływu ściągającego obrazu na lokalizowanie źródła dźwięku oraz stwierdzono, że efekt ten zachodzi niezależnie od woli osoby biorącej udział w badaniu [4] [73] [143]. Wieloletnia praca przy miksowaniu wspieranym informacją wizualną może również powodować ugruntowanie błędnych schematów kognitywnych. Jak podaje Jakubik [64]: „w sytuacjach niejednoznacznych o pewnym stopniu niepewności, człowiek może kierować się schematami poznawczymi", które „powstają na podstawie dotychczasowych doświadczeń i pozwalają kształtować oczekiwania związane z przedmiotem". Jak dalej pisze autor: „zdarza się jednak, że schematy te nie są w pełni

adekwatne do rzeczywistości i zniekształcają spostrzegany obiekt". W tym kontekście niektórzy realizatorzy krytykują ideę gotowych ustawień (ang. *presets*) w systemach DAW, zwracając uwagę na fakt, że w wielu przypadkach o wyborze danego ustawienia decyduje przyzwyczajenie zamiast faktyczna potrzeba [115]. Dodatkowo postrzeganie danej emulacji jako dobrej lub słabej, pomimo oceniania pod kątem wierności brzmienia, jest uwarunkowane estetyką interfejsu graficznego [115].

Można przypuszczać, że problem związany z podejmowaniem decyzji głównie na podstawie graficznej reprezentacji zmian wartości parametrów fonicznych dotyczy w większym stopniu niedoświadczonych realizatorów niż inżynierów miksu z wieloletnim stażem. Jednakże, w przypadku tych drugich, na podstawie ich własnych wypowiedzi [109] można stwierdzić, że problem ten również występuje. Jak wcześniej wspomniano, światowej sławy inżynierowie dźwięku twierdzą, że wynik ich pracy w oparciu jedynie o środowisko DAW jest zdecydowanie słabszy niż przy zastosowaniu stołu mikserskiego [109]. Biorąc pod uwagę różnice w obu podejściach (zgranie w oparciu o środowisko DAW i stoły mikserskie), możliwe jest postawienie hipotezy o wpływie na przebieg i wynik miksowania nie tylko ergonomii, ale również bodźców wzrokowych (wizualizacja parametrów fonicznych i ich zmian), które mogą angażować inżyniera dźwięku w zbyt dużym stopniu i zaburzać uwagę słuchową.

Można zatem zidentyfikować przypadek, w którym z jednej strony pożądane są rozwiązania uniwersalne i tanie o wysokiej ergonomii, a z drugiej istnieje potrzeba opracowania systemów pozwalających na pracę nieangażującą w znaczącym stopniu zmysłu wzroku. Rozwiązań można poszukiwać w obszarze HCI (ang. *Human-Computer Interaction*) [24]. Wiele uwagi poświęca się ostatnio rozwojowi interfejsów umożliwiających interakcję za pomocą gestów rąk. Sterowanie za pomocą gestów rąk znajduje obecnie szerokie zastosowanie, m.in. w zdalnym sterowaniu robotem [6] [138] [146] [151], w diagnostyce chorób [75] i terapii ruchowej, przeglądaniu obrazów w warunkach szpitalnych [129] [144] [145], w obsłudze gier komputerowych [84] [94], obsłudze przenośnych odtwarzaczy muzyki [68] czy wreszcie bezdotykowym przewijaniu slajdów prezentacji w trakcie prelekcji [85]. Automatyczna detekcja gestów w obrazie wizyjnym ma istotne znaczenie przy tworzeniu systemów rozpoznawania i uczenia języka migowego [9] [15] [27] [67] i miganego [1] [19] [54] [55].

Wykorzystanie swobodnych, wykonywanych w powietrzu, gestów rąk w procesie miksowania sygnałów fonicznych może pozwolić na całkowite wyeliminowanie konieczności stosowania urządzeń pośredniczących pomiędzy użytkownikiem systemu a dźwiękiem. Stworzyłoby to możliwość większego zanurzenia inżyniera miksu w procesie miksowania. Rozwiązanie takie mogłoby się również przyczynić do rozwoju nowych podejść do miksowania, kładących nacisk na aspekty artystyczne w większej mierze niż metody znane obecnie. Przykładowo, podejście takie mogłoby wykorzystywać elementy dyrygowania orkiestrą. Dyrygowanie jako proces, w ramach którego kontrolowane jest nie tylko tempo utworu, ale również dynamika i balans pod postacią zarówno różnic poziomów, jak i zależności częstotliwościowych źródeł wynikających z rozkładu rezonansów, może być uznane za rodzaj miksowania liniowego.

Dodatkową zaletą miksowania za pomocą gestów rąk rozpoznawanych w procesie analizy obrazu wizyjnego jest możliwość wyeliminowania filtracji grzebieniowej powstającej wskutek odbić dźwięku [49] od powierzchni płaskich, takich jak blat stołu lub płyta czołowa stołu mikserskiego. W takim przypadku miksowanie za pomocą gestów może stworzyć warunki, w których dźwięk pomiędzy monitorami studyjnymi a inżynierem miksu rozchodziłby się w wolnym polu. Przy zastosowaniu materiałów absorpcyjnych na podłodze, suficie i ścianach możliwe byłoby zapewnienie dźwięku pozbawionego „zakolorowań" [38] [39].

Powyższe obserwacje i spostrzeżenia stanowią **genezę tematu niniejszej rozprawy. Jako główny cel rozprawy określono opracowanie systemu miksowania dźwięku za pomocą gestów rąk wykonywanych w powietrzu oraz zbadanie możliwości oferowanych przez takie rozwiązanie w porównaniu ze współczesną metodą miksowania sygnałów fonicznych, wykorzystującą środowisko komputera.** Wśród **celów cząstkowych znalazło się m.in. sprawdzenie wpływu ergonomii na sposób i wyniki miksowania.** Jako elementy istotne z punktu widzenia ergonomii można podać wygodę obsługi, dokładność i intuicyjność interfejsu. Interesującym aspektem opracowanego systemu jest możliwość prowadzenia procesu miksowania wspieranego informacją wizualną, jak również miksu, w którym realizacja odbywa się bez bodźców wzrokowych. Zastosowanie gestów w procesie miksowania oferuje unikatową możliwość przeprowadzania tych procesów zgodnie z drugim z wyżej wymienionych sposobów, co zgodnie z sugestiami realizatorów nagrań powinno mieć

istotny wpływ na otrzymywanie zgrań o wyższych walorach estetycznych. W związku z tym jako **drugi z celów cząstkowych rozprawy określono przeprowadzenie badań wpływu obecności na ekranie informacji wizualnej odzwierciedlającej wartości parametrów sygnałów fonicznych na decyzje podejmowane w trakcie miksowania**, które warunkują wartość estetyczną zgrań. Zgodnie z definicją estetyki [44], pod pojęciem wartości estetycznych rozumieć można ogół cech nagrania ukształtowanych w procesie twórczym (w procesie miksowania), wpływających na percypowanie dzieła (zgrania) przez odbiorcę w sposób zgodny z zamierzeniami twórcy (inżyniera dźwięku). Na cechy te wpływają takie własności zgrania, jak: balans, lokalizacja źródeł w bazie stereofonicznej, dynamika, przestrzenność, klarowność, wykorzystanie efektów.

Osiągnięcie celu rozprawy wiąże się z udowodnieniem postawionych poniżej tez:

1. **Możliwe jest efektywne sterowanie procesami miksowania dźwięku za pomocą gestów interpretowanych przez komputerowy system analizy obrazu wizyjnego**
2. **Zastosowanie logiki rozmytej w procesie rozpoznawania gestów dynamicznych, dla których trajektorią ruchu jest okrąg, pozwala na ich interpretację z wysoką skutecznością.**

W związku z wcześniej przedstawionymi rozważaniami nt. wpływu informacji wizualnej wyświetlanej na ekranie monitora na proces miksowania w systemach DAW, w kolejnym rozdziale przedstawiono wybrane zagadnienia związane z jednoczesną percepcją dźwięku i obrazu.

2 Percepcja i interakcja wielomodalna

Wiele badań wskazuje, że zmysł wzroku pełni wiodącą rolę w percepcji otaczającej rzeczywistości [4] [70] [76] [143]. Za przykład może posłużyć powszechne zjawisko ściągającego wpływu obrazu na lokalizację źródła dźwięku, podczas gdy zjawisko odwrotne zachodzi stosunkowo rzadko [4] [73] [143]. Ze względu na fakt, że modalność wzrokowa przeważa w percepcji bodźców, jednoczesne angażowanie zmysłu wzroku i słuchu może spowodować ograniczenie roli modalności słuchowej [13] [143]. W ogólności, zgodnie z hipotezą Welcha i Warrena, stopień zaangażowania danej modalności w procesie percepcji zależy od typu analizowanej cechy i rodzaju aktywności [4]. Powszechnie wiadomo, że trwałe wyłączenie jednego z analizatorów z procesów percepcji, np. na skutek kalectwa, powoduje obniżenie się progów wrażliwości sensorycznej na bodźce w pozostałych analizatorach [13] [76]. Natomiast, jak podaje Bogdanowicz, powołując się na prace Włodarskiego: „stymulowanie jednych narządów zmysłu powoduje różnorodne zmiany w funkcjonowaniu innych, np. pod wpływem bodźców dźwiękowych zmieniają się progi wrażliwości i czułości wzrokowej, a określone oświetlenie sprawia, że dźwięki są lepiej słyszane i wydają się głośniejsze" [13] [148]. Z tych względów, podstawowym założeniem systemu było zapewnienie możliwości wykonywania wszystkich funkcji związanych bezpośrednio z procesem miksowania nagrania muzycznego za pomocą gestów przy braku informacji wizualnej. Obecnie nie są znane systemy komputerowe, umożliwiające bez zastosowania zaawansowanego kontrolera zastępującego stół mikserski, produkcję dźwięku bez angażowania w znaczącym stopniu zmysłu wzroku. Wydaje się, iż głównym powodem takiego stanu rzeczy są ograniczenia tradycyjnych i powszechnie dostępnych interfejsów: myszy i klawiatury, wykorzystywanych w profesjonalnych aplikacjach do pracy z dźwiękiem. Specyfika pracy z tymi interfejsami wpłynęła na wypracowanie pewnego standardu prezentowania i zarządzania informacją w aplikacjach do produkcji dźwięku. Standard ten, chociaż umożliwia relatywnie efektywne pod względem czasowym wykonywanie procesów obróbki dźwięku w porównaniu z klasycznym sposobem miksowania dźwięku z wykorzystaniem stołu mikserskiego i zewnętrznych urządzeń przetwarzających dźwięk, jest często krytykowany przez profesjonalnych realizatorów nagrań. Przedmiotem krytyki jest między innymi wspomniana niemożność wyłączenia wrażeń wizualnych z procesu

miksowania nagrania, mogąca powodować zaburzenia bądź ograniczenie percepcji dźwięku. Potwierdzeniem występowania takich zaburzeń może być efekt McGurka [48], zgodnie z którym na podstawie bodźców kierowanych jednocześnie do dwóch różnych zmysłów wypracowana zostaje przez system aferentny odpowiedź niezgodna z żadnym z pobudzeń. Przykładowo, osoba której na ekranie monitora przedstawiono obraz twarzy spikera wypowiadającego sylabę *ga*, ale jako bodziec akustyczny podano sylabę *ba*, w efekcie odpowiada, że spercypowała sylabę *da*. W odniesieniu do oprogramowania miksowania dźwięku można zauważyć, że ze względu na silnie zróżnicowany charakter przetwarzanych dźwięków, nie jest możliwe jednoznaczne odzwierciedlenie zmiany danego parametru dźwiękowego w postaci zmiany w wyglądzie elementu graficznego. Przykładowo, niewielka różnica w ustawieniu wirtualnego pokrętła modyfikującego parametr może powodować znaczącą zmianę dźwięku i odwrotnie – istotna zmiana w wyglądzie tego samego elementu graficznego może powodować ledwie zauważalną zmianę cech dźwięku. W efekcie percypowany dźwięk może być wypadkową pobudzenia wizualnego i akustycznego. Dodatkowo wspomniany standard prezentowania informacji użytkownikowi może prowadzić do przywoływania z pamięci i powielania wcześniej wykorzystanych wzorców myślowych związanych z graficznym wyglądem elementu reprezentującego dany parametr lub zbiór parametrów zamiast do podejmowania akcji faktycznie adekwatnych do percypowanego materiału dźwiękowego. Ma to związek ze wspomnianym we wprowadzeniu problemem kierowania się schematami poznawczymi, które mogą być nieadekwatne do rzeczywistości i zniekształcać postrzegany obiekt [64]. W literaturze poświęconej zagadnieniu produkcji muzycznej opis tego zjawiska tłumaczy się również jako podświadome przekonanie o tym, że to co „wizualnie prezentuje się lepiej – brzmi lepiej". Fakt ten wykorzystują producenci programowych wtyczek przetwarzających dźwięk starając się, aby wygląd interfejsu graficznego jak najlepiej odzwierciedlał oczekiwane efekty brzmieniowe. Biorąc pod uwagę te spostrzeżenia i możliwości związane z zastosowaniem interfejsu HCI (ang. *Human-Computer Interaction*) wykorzystującego gesty rąk, istotne wydaje się spełnienie postawionego wymagania dotyczącego obsługi procesów miksowania dźwięku bez angażowania zmysłu wzroku.

Dodatkowym ograniczeniem systemów produkcji dźwięku obsługiwanych jedynie za pomocą myszy i klawiatury jest brak zapewnienia możliwości jednoczesnej edycji

więcej niż jednego parametru. W istocie jest to ograniczenie znaczące, gdyż zmiana jednego parametru charakteryzującego dźwięk może wpływać na percepcję innego, kształtując w niezamierzony sposób odbiór całości. Opracowany system wychodzi naprzeciw temu problemowi poprzez udostępnianie możliwości edycji jednocześnie więcej niż jednego parametru za pomocą gestów obu rąk użytkownika.

Przegląd literatury w poszukiwaniu badań dotyczących wpływu percepcji wielomodalnej i sposobu interakcji na wyniki procesu miksowania pozwala stwierdzić, że problemy przedstawione w niniejszej rozprawie mają charakter nowatorski. Nie natrafiono bowiem na publikacje, których istotą byłoby zbadanie zależności pomiędzy wynikami miksowania a specyfiką wielomodalnej percepcji i interakcji typowej dla systemów DAW. Jak wspomniano we Wprowadzeniu, problem ten jest jednak często poruszany przez znanych inżynierów dźwięku [22] [32] [43] [90] [91] [109]. Dlatego ważne jest, że proponowany w rozprawie system, wykorzystujący w procesie miksowania interakcję za pomocą gestów rąk, oferuje możliwość zbadania tych aspektów. Z tego względu w kolejnym rozdziale przedstawiono wybrane metody rozpoznawania dynamicznych i statycznych gestów rąk. Metody te wybrano ze względu na ich wykorzystanie w systemie umożliwiającym efektywną obsługę procesów miksowania dźwięku.

3 Wybrane metody rozpoznawania gestów rąk

Definicyjnie gesty rąk można określić jako posiadający znaczenie statyczny lub zmienny w czasie układ rąk i dłoni, który może być wykorzystany przy interakcji ze środowiskiem [96]. Metody rozpoznawania gestów leżą w kręgu zainteresowań obszaru HCI. Jak wspomniano wcześniej we Wprowadzeniu wśród licznych dziedzin, w których rozpoznawanie gestów rąk znajduje zastosowanie można wymienić: robotykę, multimedia czy medycynę. W dziedzinach tych istotnym problemem warunkującym skuteczne rozpoznawanie gestów jest wstępne przetwarzanie obrazu wizyjnego. W literaturze poświęconej problemowi rozpoznawania gestów w obrazie wizyjnym można znaleźć podstawy teoretyczne zarówno metod przetwarzania obrazu, jak i śledzenia ruchów rąk czy detekcji kształtów dłoni [23] [29] [40] [53] [94] [96] [112] [114] [129] [141]. Zagadnienie wstępnego przetwarzania obrazu, stanowiące pewien wydzielony obszar opracowanego systemu rozpoznawania gestów, zostało dokładnie opisane w rozdziale 5.8.1. Bloki przetwarzania obrazu zastosowane w systemie przygotowanym w ramach rozprawy stanowią przykład typowych operacji przetwarzania stosowanych w podejściach do rozpoznawania gestów. Dodatkowo metoda przetwarzania obrazu została rozszerzona o autorskie pomysły, wynikające z przyjętych założeń dotyczących komponentów systemu. W niniejszym rozdziale zaprezentowano w pierwszej kolejności rozwiązania systemowe stosowane w rozpoznawaniu gestów, a następnie w tym kontekście przedstawiono wybrane metody rozpoznawania gestów statycznych i dynamicznych, które leżą u podstaw tych rozwiązań.

W ogólności gesty rąk podzielić można na statyczne, tj. gesty, których istotę stanowi kształt dłoni, i dynamiczne, tj. takie, dla których istotna informacja zawarta jest w trajektorii ruchu przedramienia, dłoni bądź palców [96]. Możliwe jest jednoczesne rozpoznawanie gestów statycznych i dynamicznych i przypisywanie znaczeń gestom będącym kombinacją obu rodzajów [96].

W oparciu o przegląd badań w dziedzinie rozpoznawania gestów wyróżnić można trzy podstawowe sposoby pozyskiwania informacji o geście. Pierwszy ze sposobów polega na wykorzystaniu zakładanej na dłoń rękawiczki wyposażonej w czujniki [8] [96] [133]. Rozwiązanie to zapewnia wysoką skuteczność detekcji gestów ze względu na stosunkowo łatwą możliwość pozyskiwania z czujników dokładnych danych reprezen-

tujących ruch i kształt dłoni. Dodatkowo, przy zastosowaniu rękawiczki zawierającej czujniki w liczbie pozwalającej na śledzenie zmian kątów pomiędzy wszystkimi kośćmi palców, możliwe jest rozpoznawanie wszelkich kształtów, w jakie człowiek jest w stanie uformować dłoń [98] [100]. Wadą systemów opartych na zastosowaniu rękawiczki jest jednak ograniczenie swobody użytkownika, szczególnie w przypadku stosowania rękawiczek połączonych przewodem z urządzeniem odbierającym dane (komputerem). Jedną z metod zwiększenia swobody ruchów użytkownika jest zaprojektowanie systemu zgodnie z drugim ze sposobów pozyskiwania informacji o geście. W sposobie tym, zamiast rękawiczki, na dłoni znajdują się znaczniki śledzone w strumieniu wizyjnym pozyskanym z kamery [35]. Pozycje znaczników oraz relacje między nimi dostarczają informacji, na podstawie których można wnioskować o wykonywanych gestach dynamicznych i statycznych. Sposób ten charakteryzuje się większą ergonomią i niższym kosztem, dzięki wyeliminowaniu stosowania drogich czujników, jednak ze względu na konieczność każdorazowego naklejania znaczników na palce, przygotowanie do użycia systemu może być postrzegane jako uciążliwe. Z tego powodu obecnie wiele uwagi poświęca się systemom zaprojektowanym zgodnie z trzecim ze sposobów pozyskiwania informacji o gestach, tj. z wykorzystaniem jedynie analizy obrazu wizyjnego. W ramach tego sposobu również wyróżnić można dwa główne nurty prac. W pierwszym nurcie proces detekcji rąk i rozpoznawania gestów wspomagany jest przez wykorzystanie emiterów i czujników podczerwieni [23] [29] [40] [53] [94] oraz kamer ToF (ang. *Time-of-Flight*) [112] [114] [129] [141] udostępniających informację o odległości od obiektu (głębi). Informacja o głębi może być również pozyskiwana z wykorzystaniem dwóch kamer [128]. Ambicją twórców prac z drugiego nurtu jest tworzenie systemów opartych wyłącznie na prostej, powszechnie dostępnej kamerze internetowej podłączanej do portu USB komputera, działającej w zakresie pasma widzialnego. Niezależnie od zastosowania zwykłych kamer bądź kamer pracujących w zakresie pasma światła podczerwonego, dąży się do tego, aby oba systemy nie nakładały na użytkownika ograniczeń w zakresie koloru skóry, ubioru, oświetlenia czy tła [96]. Rozpoznawanie gestów przy użyciu kamery USB w takich warunkach nie jest zagadnieniem trywialnym, w związku z czym, pomimo opracowania szeregu metod przetwarzania obrazu i klasyfikacji gestów, istnieje potrzeba poszukiwania nowych rozwiązań algorytmicznych. Po-

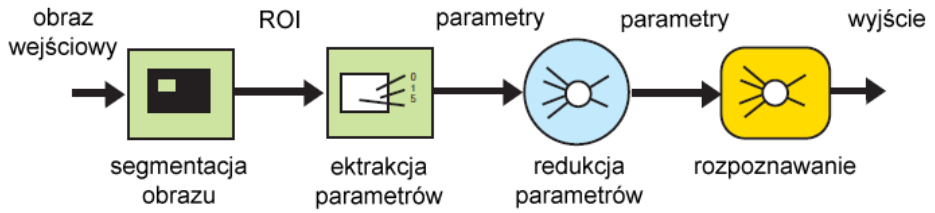
nieważ system przedstawiony w rozprawie oparto na typowej kamerze podłączanej do portu USB, w niniejszym rozdziale skupiono się na scharakteryzowaniu metod wykorzystujących kamerę tego typu.

Ze względu na obszerność praktycznej części rozprawy, wynikającą ze złożoności poruszonych w niej zagadnień, w niniejszym rozdziale ograniczono się do przedstawienia jedynie istotnych aspektów popularnych metod stosowanych przy rozpoznawaniu gestów. Metody te przedstawiono w kontekście problemu rozpoznawania gestów bez szczegółowego przytaczania leżących u ich podstaw teorii, pozwalających na ich zastosowanie przy rozwiązywaniu także innych problemów. Podstawy teoretyczne są szeroko opisane w literaturze, dlatego autor rozprawy przywołuje te prace w ramach przedstawiania każdej z metod.

3.1 Wybrane metody rozpoznawania gestów statycznych

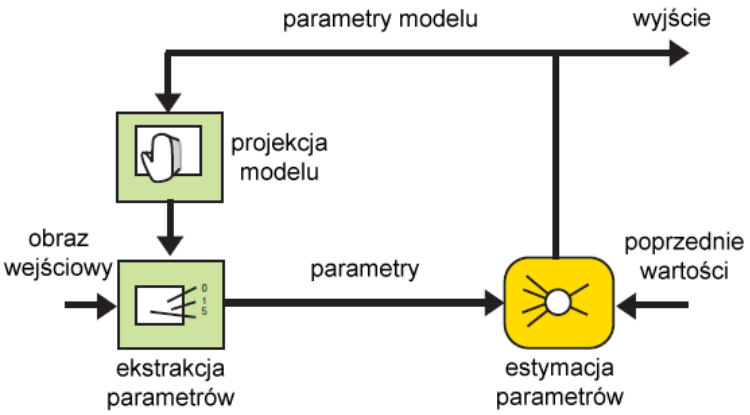
Problem rozpoznawania gestów statycznych w ogólności polega na zidentyfikowaniu statycznego układu części ciała (ręka – dłoń) i sprowadza się do ekstrakcji i klasyfikacji wektorów cech opisujących gest. Aby identyfikacja kształtu dłoni była możliwa, z reguły konieczne jest najpierw odpowiednie przetworzenie obrazu. Jak wspomniano na początku rozdziału 3. operacje przetwarzania obrazu zostały wyszczególnione w rozdziale 5.8.1.

Metody rozpoznawania gestów statycznych można podzielić pod względem architektury na dwie podstawowe kategorie: bazujące na danych i bazujące na modelach [96] [142]. Systemy bazujące na danych na podstawie obserwacji obiektu w obrazie pozyskanym z kamery dokonują bezpośredniego mapowania kształtu na jedną z klas gestów. W tym celu z przetworzonego obrazu ekstrahowany jest fragment zawierający dłoń, który jest następnie analizowany w kontekście wydobycia cech reprezentujących jej kształt (rys. 3.1). Pozyskane cechy podawane są na wejście wybranego klasyfikatora w postaci wektora cech. Metody ekstrakcji dłoni z obrazu i jej parametryzacji przedstawiono w dalszej części niniejszego rozdziału.



Rys. 3.1 Architektura systemu rozpoznawania gestów statycznych bazująca na danych [142]

Systemy drugiej z wyszczególnionych kategorii wykorzystują model dłoni, który stanowi reprezentację różnych, dozwolonych kombinacji kształtów klasyfikowanych jako ten sam gest. Ze względu na określenie wielu kształtów za pomocą jednego modelu, konieczne jest przeprowadzenie procedury dopasowania cech obiektu znajdującego się w obrazie do stosownego modelu (rys. 3.2). Na podstawie tego dopasowania otrzymywana jest bezpośrednio informacja o wykonanym geście.



Rys. 3.2 Architektura systemu rozpoznawania gestów statycznych bazująca na modelach [142]

Wśród modeli można również wyróżnić dwie podstawowe kategorie: modele prezencji (ang. *appearance based models*) i modele fizyczne. Modele prezencji definiują cechy dłoni widzianej pod różnymi postaciami w obrazie dwuwymiarowym [142]. Modele fizyczne odzwierciedlają rzeczywisty kształt dłoni, przez co wymagają użycia procedury mapowania parametrów pomiędzy przestrzeniami: dwuwymiarową i trójwymiarową [142].

3.1.1 Metody ekstrakcji obszaru dłoni z obrazu

Aby parametryzacja kształtu dłoni i w dalszej kolejności klasyfikacja gestu w systemach o architekturze opartej na danych były możliwe, konieczne jest najpierw wyodrębnienie dłoni z obrazu. Proces ten ma znaczenie zarówno z punktu widzenia rozpoznawania gestów statycznych, jak i dynamicznych, chociaż w przypadku tych drugich znane są metody pozwalające na wykrywanie przemieszczenia dłoni i wykreślanie trajektorii ruchu bezpośrednio w nieprzetworzonym obrazie z kamery [17]. Skuteczność wyodrębnienia dłoni z obrazu zdeterminowana jest przez dobór metody przetwarzania obrazu, odpowiedniej dla warunków, w jakich ma pracować system. Warunki te mogą dotyczyć koloru dłoni użytkownika, wpływu oświetlenia na równomierność i kolor zabarwienia dłoni, koloru i zmienności tła za użytkownikiem, wielkości dłoni w obrazie, jak również złożoności słownika gestów. W najprostszych podejściach [28] [34] [50] [60] stosuje się progowanie obrazu, przyjmując założenie, że kontrast pomiędzy ręką a pozostałymi elementami obrazu jest zawsze wystarczający do jej niezawodnego wyekstrahowania. W rozwiązaniach, w których kamera umieszczona jest przed użytkownikiem i skierowana na niego, jedną z metod spełnienia tego warunku jest oświetlanie wyciągniętej przed siebie dłoni dodatkowym światłem. Inną metodą, która w połączeniu z podaną powyżej może dodatkowo podnieść skuteczność ekstrakcji, jest zastosowanie rękawiczki w unikatowym kolorze [9]. Rozwiązaniem bardziej zaawansowanym od progowania jest wykorzystanie modelu koloru skóry [130]. Dla warunków, w których w analizowanym obrazie nie pojawia się twarz użytkownika, a tło za użytkownikiem ma inny kolor niż dłoń, metoda zapewnia wysoką skuteczność ekstrakcji. Rozwiązaniem problemu obecności w kadrze poza rękoma również twarzy może być zastosowanie do jej detekcji kaskady Haara [37]. Powtarzalny rozkład oczu, nosa i ust pozwala za pomocą tej metody skutecznie wykrywać twarz w obrazie i w ten sposób wykluczyć zawierający ją fragment z procesu ekstrakcji rąk. W dalszym ciągu jednak problem stanowi wyodrębnienie kształtu dłoni w sytuacji, gdy znajduje się ona na tle twarzy.

W systemach o architekturze opartej na modelach ekstrakcja dłoni polega na przeszukiwaniu obrazu metodą jego okienkowania ze stałym krokiem [17] [142]. Fragment obrazu wyznaczony przez wielkość i lokalizację okna analizowany jest pod kątem

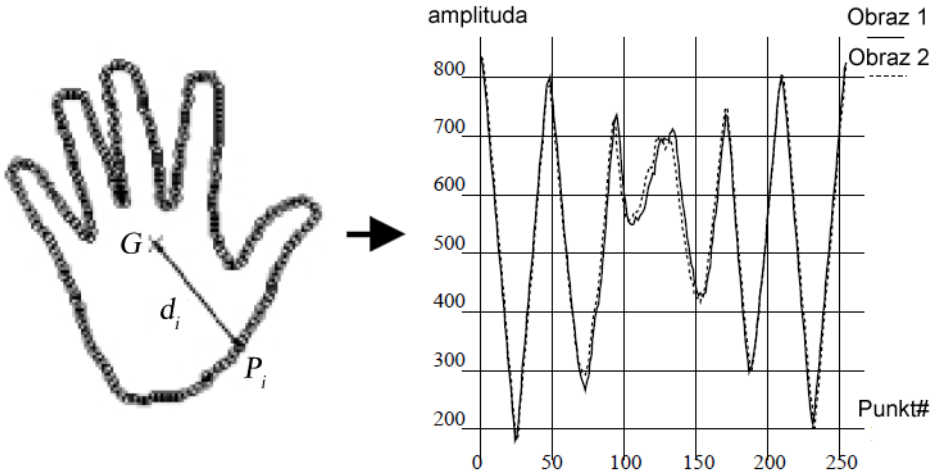
dopasowania do wzorca. Detekcja dłoni, realizowana poprzez dopasowanie do modelu, zapewnia jednocześnie informację o wykrytym geście statycznym.

Ponieważ w systemie opracowanym w ramach niniejszej rozprawy, kamera jest umieszczona przed użytkownikiem i skierowana nie na niego, tylko na ekran, nie jest możliwe wykorzystanie metody bazującej na modelu koloru skóry. Obraz pozyskiwany przez kamerę zawiera zniekształcony obraz wyświetlany przez projektor, zatem jako optymalną wybrano metodę ekstrakcji dłoni, polegającą na progowaniu odpowiednio przetworzonej różnicy tych obrazów. Metoda ta została szczegółowo opisana w rozdziałach 5.8.1 i 5.8.2.

3.1.2 Metody parametryzacji dłoni

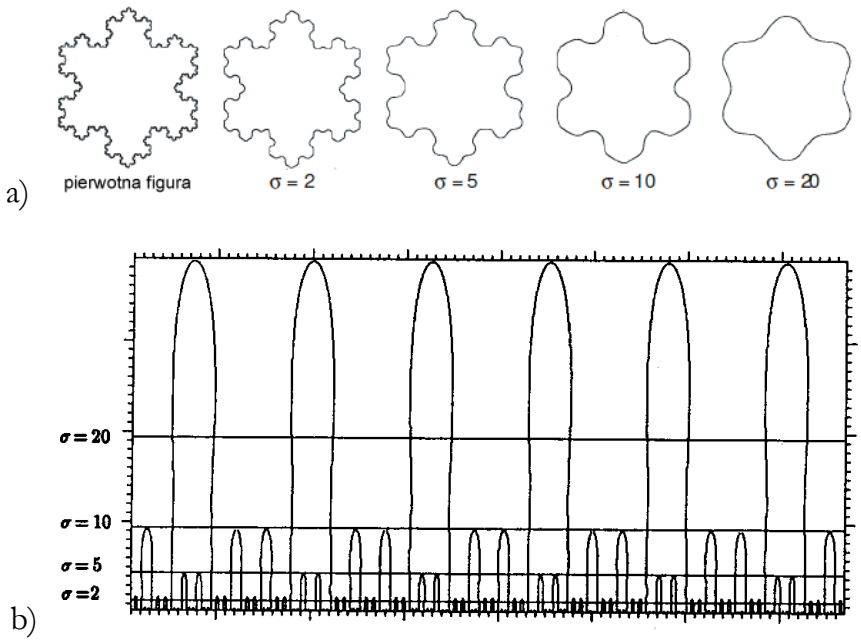
W praktyce stosowane są różne metody pozyskiwania cech (parametryzacji) dłoni, które w zależności od warunków, w jakich ma pracować system, mogą zapewniać niewrażliwość na zmiany oświetlenia, skalę, obrót lub translację.

Jedna z powszechnych metod parametryzacji oparta jest na wyznaczaniu punktów na konturze kształtu otrzymanego w procesie progowania i obliczaniu odległości każdego z nich od pewnego punktu charakterystycznego. Przykładowo Hamada i in. wyznaczają 256 punktów leżących na konturze dłoni widocznej w obrazie i następnie obliczają odległość każdego z nich od środka ciężkości kształtu ograniczonego konturem [50]. Uzyskane odległości umieszczane są w wektorze parametrów w taki sposób, że ich wykres (rys. 3.3) rozpoczyna się od maksymalnego skoku. W ten sposób uzyskuje się uniezależnienie parametrów od obrotu dłoni względem osi przechodzącej przez płaszczyznę ekranu. Odległości są dodatkowo normalizowane, co zapewnia niewrażliwość na skalę. Ponieważ odległości wyznaczane są w stosunku do środka ciężkości, metoda cechuje się dodatkowo niewrażliwością na translację dłoni w obrazie.



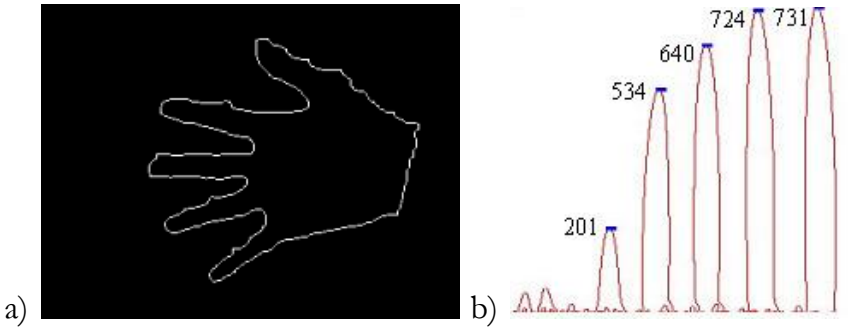
Rys. 3.3 Metoda parametryzacji kształtu dłoni bazująca na punktach (P_i) leżących na konturze dłoni i ich odległościach (d_i) od środka ciężkości (G_i) [50]

Inną metodą wykorzystującą kontur dłoni do opisu jej cech jest metoda przestrzeni krzywizn skalowalnych (ang. *Curvature Scale Space;* CSS) [101]. Metoda ta, podobnie jak przedstawiona powyżej, również uniezależnia rozkład parametrów od skali i translacji dłoni w obrazie oraz od jej obrotu względem osi przechodzącej przez płaszczyznę ekranu. Przestrzeń krzywizn skalowalnych jest zbiorem różnych reprezentacji krzywizny konturu obiektu, różniących się od siebie poziomem szczegółowości. Zanižanie stopnia szczegółowości osiągane jest przez splot funkcji definiującej pierwotną krzywą z funkcją Gaussa. Dla zilustrowania metody, na rys. 3.4 przedstawiono krzywą Kocha wraz z jej kolejnymi przybliżeniami oraz reprezentację otrzymanej przestrzeni krzywych w postaci obrazu CSS.



Rys. 3.4 Krzywa Kocha i jej kolejne przybliżenia (a) oraz obraz CSS dla krzywych (b) [101]

Kolejne krzywe otrzymywane przez zaniżanie stopnia szczegółowości odzwierciedlane są w obrazie CSS przez coraz większe piki, o mniejszej liczbie przejść przez zero. Przykładowy obraz CSS dla konturu dłoni przedstawiono na rys. 3.5.



Rys. 3.5 Przykładowy obraz CSS (b) reprezentujący kontur dłoni (a) [27]

Chang i Pengwu stosują do parametryzacji kształtu dłoni zredukowaną formę obrazu CSS, przechowywaną w *n*-elementowym wektorze o stałym rozmiarze [27]. Każdy element w wektorze reprezentuje jeden z wyznaczonych ze stałym rozmiarem okna fragmentów obrazu CSS. Element ten przyjmuje wartość maksymalną najbardziej znaczącego szczytu zlokalizowanego we fragmencie obrazu. Taka modyfikacja klasycznej metody CSS pozwoliła na bezpośrednie wykorzystanie powstałego wektora parame-

trów jako wektora stanu dla ukrytych modeli Markowa wykorzystanych przez autorów wspomnianej pracy do rozpoznawania gestów.

3.1.3 Metody rozpoznawania statycznych gestów dłoni oparte na modelach

Zgodnie z informacjami zawartymi na początku rozdziału 3.1 zastosowanie systemu rozpoznawania gestów o architekturze opartej na modelach prezencji lub modelach fizycznych pozwala na pozyskanie informacji o wykonanym geście dłoni bezpośrednio w procesie dopasowania do modelu obiektu znajdującego się we fragmencie obrazu.

Modele prezencji

Modele prezencji stanowią reprezentację statystyczną wyglądu obiektu pozwalającą na wyodrębnienie go z obrazu, przy czym jego nieznaczne deformacje oraz zmiana kąta obserwacji nie wpływają na skuteczność wyodrębniania. W tym kontekście, za powszechną uznać można metodę *aktywnych modeli kształtu* [52] [62]. W metodzie tej w pierwszej kolejności na kontury dłoni w obrazach wzorców w zbiorze treningowym nanoszone są ręcznie punkty kluczowe. Istotne jest, aby punkty rozmieszczone były w sposób jednolity, tzn. oznaczenie wybranego miejsca na konturze dłoni powinno odpowiadać oznaczeniu tego samego punktu w kolejnym obrazie. Następnie za pomocą operacji skalowania, obrotu i translacji zmieniane jest położenie punktów oraz określany jest kształt uśredniony. Istotą operacji jest zminimalizowanie sumy kwadratów odległości pomiędzy punktami każdego kształtu a punktami kształtu uśrednionego. Kształty te, w ramach zbioru treningowego są redukowane, np. z wykorzystaniem metody PCA (ang. *Principal Component Analysis*), w celu otrzymania najbardziej znaczących reprezentacji. Istotą metody jest takie ograniczenie zbioru danych, aby ich wariancja była maksymalna. Zastosowanie metody FDA (ang. *Fisher Discriminant Analysis*) zamiast metody PCA, zgodnie z doniesieniami literaturowymi [46] [47], pozwala zwiększyć skuteczność rozpoznawania kształtów. Metoda ta w odróżnieniu od metody PCA dokonuje przekształceń, które prowadzą do uzyskania danych najważniejszych z punktu widzenia efektywnego rozpoznawania, a nie jedynie danych o największej wariancji. W

oparciu o parametry tych kształtów oraz kształt uśredniony, definiowany jest model rozmieszczenia punktów – PDM (ang. *Point Distribution Model*). Proces dopasowywania modelu do kształtu w obrazie polega na modyfikacji tych parametrów w ramach granic wytyczonych przez przykłady ze zbioru treningowego. W ten sposób możliwe jest dopasowanie modelu do obecnego w obrazie kształtu, nieokreślonego wcześniej w zbiorze treningowym.

Kształt dłoni można również modelować za pomocą grafów elastycznych [139]. W metodzie tej na obrazie dłoni rozmieszczane są łączone odcinkami punkty (rys. 3.6). Długości odcinków określają dozwolone odległości pomiędzy punktami. W odróżnieniu od metody *aktywnych modeli kształtu* zastosowanej do tworzenia reprezentacji kształtu dłoni, w metodzie grafów elastycznych punkty wyznaczane są nie tylko na konturze dłoni, ale w obrębie całego kształtu.

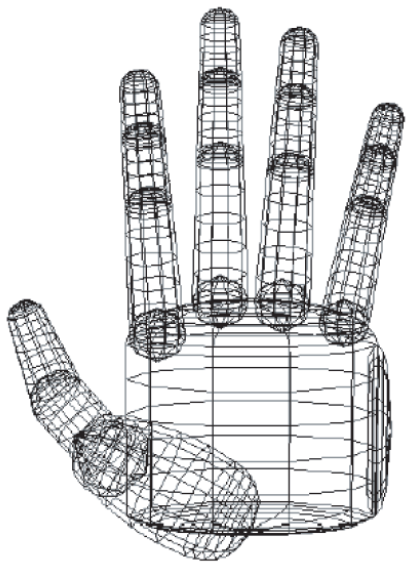
Rys. 3.6 Przykładowy kształt dłoni zamodelowany za pomocą grafów elastycznych [139]

Modele prezencji mogą być również zastosowane do reprezentacji tekstury modelowanego obiektu. W tym celu możliwe jest wykorzystanie uogólnienia metody *aktywnych modeli kształtu*, tj. *aktywnych modeli prezencji* [41]. Parametry opisujące kształt są łączone z parametrami definiującymi teksturę i w postaci takiego zbioru poddawane redukcji z wykorzystaniem metody PCA.

Modele fizyczne

Modele fizyczne najczęściej występują w postaci szkieletu lub połączonych płaszczyzn przedstawiających dłoń uformowaną w określony kształt (rys. 3.7). Kształt ten

może być zamodelowany w programie graficznym umożliwiającym tworzenie grafiki pseudo-trójwymiarowej, takim jak np. Autodesk 3ds Max (dawniej: 3d Studio), Blender. Wu i in. zastosowali podejście, w którym generują modele fizyczne na podstawie danych pozyskanych z rękawiczki wyposażonej w czujniki [149]. Pierwotne dane są dwukrotnie redukowane – najpierw do siedmiu wymiarów za pomocą metody PCA, a następnie do 28 podstawowych konfiguracji. Pozostałe konfiguracje kształtów dłoni są generowane w oparciu o liniową kombinację wybranych dwóch konfiguracji podstawowych. Za pomocą modeli fizycznych, poza samym kształtem, można również modelować dynamikę ruchów palców i dłoni. Autorzy pracy oparli modelowanie dynamiki na prostym procesie stochastycznym, zwanym błądzeniem losowym [149]. W innym podejściu [132] zastosowano proces stacjonarny drugiego rzędu [31].



Rys. 3.7 Przykład modelu trójwymiarowego ludzkiej dłoni [132]

3.1.4 Metody klasyfikacji gestów dłoni

Klasyfikacja gestów dłoni jest końcowym etapem przetwarzania w systemach o architekturze opartej na danych. W najprostszym ujęciu klasyfikację można przedstawić jako problem dopasowania wzorcowego zbioru wektorów parametrów do zbioru wektorów parametrów wejściowych reprezentujących pewne przybliżenie kształtu wzorcowego. Znanych jest wiele metod klasyfikacji, spośród których jako najczęściej stosowane można wymienić: metodę k-najbliższych sąsiadów [18] [72], sztuczne sieci neu-

ronowe [98], maszyny wektorów nośnych [26] [61], drzewo C4.5 [11] [117], naiwną sieć Bayesa [56] [124] [150] czy drzewa i lasy losowe [147].

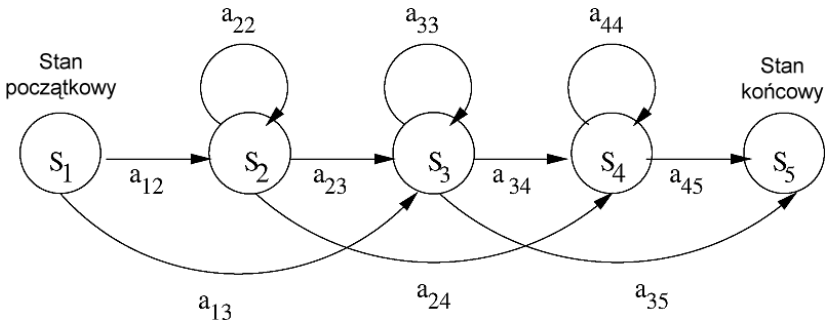
Wspomniane klasyfikatory zbadano w celu określenia optymalnego klasyfikatora dla systemu, a wyniki badań przeprowadzonych przez Autora przedstawiono w rozdziale 5.2. W tym samym rozdziale przedstawione zostały parametry klasyfikatorów.

3.2 Wybrane metody rozpoznawania gestów dynamicznych

Podstawowym problemem związanym z efektywnym rozpoznawaniem gestów dynamicznych jest wyodrębnienie gestu z sekwencji ruchów, mającego określone znaczenie. Zadanie to nie jest trywialne ze względu na niejednoznaczność segmentacji i zróżnicowanie czasowo-przestrzenne. Istotą pierwszego z wymienionych aspektów jest występowanie wraz z ruchami będącymi przedmiotem zainteresowania ruchów, które nie posiadają znaczenia przypisanego gestom. Ruchy nie posiadające znaczenia mogą występować w momentach przejść pomiędzy gestami lub wynikać z wygody użytkownika. W systemie charakteryzującym się pełną automatyzacją i efektywnością detekcji gestów stwarza to trudności związane bezpośrednio z koniecznością zidentyfikowania momentów rozpoczęcia i zakończenia gestu. Zróżnicowanie czasowo-przestrzenne związane jest z faktem, że ten sam gest może być za każdym razem wykonywany w odmienny sposób, nawet przez tego samego użytkownika systemu. Dotyczy to zarówno czasu trwania określonego gestu, trajektorii ruchu, jak i kształtu ręki widzianej w obrazie z kamery. Z tego względu w większości zaawansowanych rozwiązań przykłada się dużą wagę do modelowania stanów, za pomocą których wyrazić można gest. Spośród najpopularniejszych metod realizujących to zadanie wymienić można ukryte modele Markowa [119], automaty skończone [58] czy metodę dynamicznego marszczenia czasu [89] [102]. Metody te należą do grupy metod modelowania matematycznego. W innym podejściu do określenia zależności czasowo-przestrzennych stosuje się metody inteligencji obliczeniowej [95] [108] [122] [125], a w szczególności metody logiki rozmytej [153]. Eliminację lokalnych błędów w detekcji punktów tworzących trajektorię ruchu realizuje się często w oparciu o filtry Kalmana [69] [71].

3.2.1 Modelowanie gestów za pomocą stanów

Najczęściej przytaczaną metodą w kontekście modelowania gestów dynamicznych są ukryte modele Markowa [28] [96] [119]. Ukryty model Markowa to proces stochastyczny określony przez łańcuch Markowa o skończonej liczbie stanów i zbiór funkcji losowych, z których każda powiązana jest z jednym stanem. W każdej, dyskretnej chwili czasu proces jest w jednym ze stanów i generuje obserwację, zgodną z funkcją losową odpowiadającą aktualnemu stanowi. Wyróżnić można dwa podstawowe typy ukrytych modeli Markowa: model ergodyczny oraz model Bakisa (rys. 3.8). Pierwszy z nich zakłada pełną dowolność przejść pomiędzy stanami. W drugim, kolejne stany osiągane są zgodnie z porządkiem wymuszonym liniowym upływem czasu, tj. nie jest możliwy powrót do poprzedniego stanu, stąd model ten określa się również mianem modelu „od lewej do prawej". W zagadnieniu rozpoznawania gestów dynamicznych stosuje się drugi z przytoczonych modeli. Dla każdego gestu tworzony jest osobny model Markowa. Modele te łączone są równolegle w sieć. Istotnym problemem jest określenie optymalnej struktury każdego z modelu tworzącego sieć. Przykładowo, w zagadnieniu rozpoznawania gestów języka migowego stany modelu mogą reprezentować poszczególne wizemy (wizyjne odpowiedniki fonemów) [15] [130]. Wizemy reprezentowane są przez dające się wyodrębnić trajektorie ruchów tworzących gesty. O skuteczności ukrytych modeli Markowa może świadczyć fakt, że już pod koniec lat 90. ubiegłego wieku, pomimo znacząco słabszych jednostek obliczeniowych niż obecnie, narzędzie to pozwalało na uzyskanie skuteczności na poziomie 91,3% w systemie Starnera, rozpoznającym 40 słów amerykańskiego języka migowego [130], 90% w systemie Lianga, rozpoznającym 250 słów tajwańskiego języka migowego [88] i 94,3% w systemie rozpoznającym 131 słów koreańskiego języka migowego [111]. Ostatni z przytoczonych systemów bazuje na 14 gestach dynamicznych, 23 gestach statycznych i 14 orientacjach dłoni w przestrzeni.



Rys. 3.8 Pięciostanowy ukryty model Markowa typu Bakisa [49]

W podobny sposób, jak za pomocą ukrytych modeli Markowa, gesty dynamiczne można również modelować za pomocą automatów skończonych [59]. Metoda ta modeluje zachowanie systemu dynamicznego za pomocą tablicy dyskretnych przejść pomiędzy kolejnymi stanami. Podobnie, jak w metodzie ukrytych modeli Markowa, pojedynczemu gestowi może odpowiadać jeden automat skończony. W automacie takim stany definiują określony ruch, rozpatrywany najczęściej w ujęciu kierunku przemieszczenia. Przykładowo Yeasin i Chaudhuri za pomocą automatów skończonych zamodelowali 5 gestów dynamicznych odpowiadających komendom symbolicznym, takim jak: *podejdź bliżej*, *odejdź dalej*, *przesuń się w lewo*, *przesuń się w prawo*, *zatrzymanie awaryjne* (rys. 3.9) [151]. Pętle własne modelują moment zatrzymania ręki występujący bezpośrednio przed zmianą kierunku ruchu. Etykiety numeryczne o wartości 1 reprezentują ruch (0 – brak ruchu). Przykładowa sekwencja przejść stanów dla gestu *przesuń się w prawo* przyjmuje postać S-L-R-R-L-L-R-L-R-L.

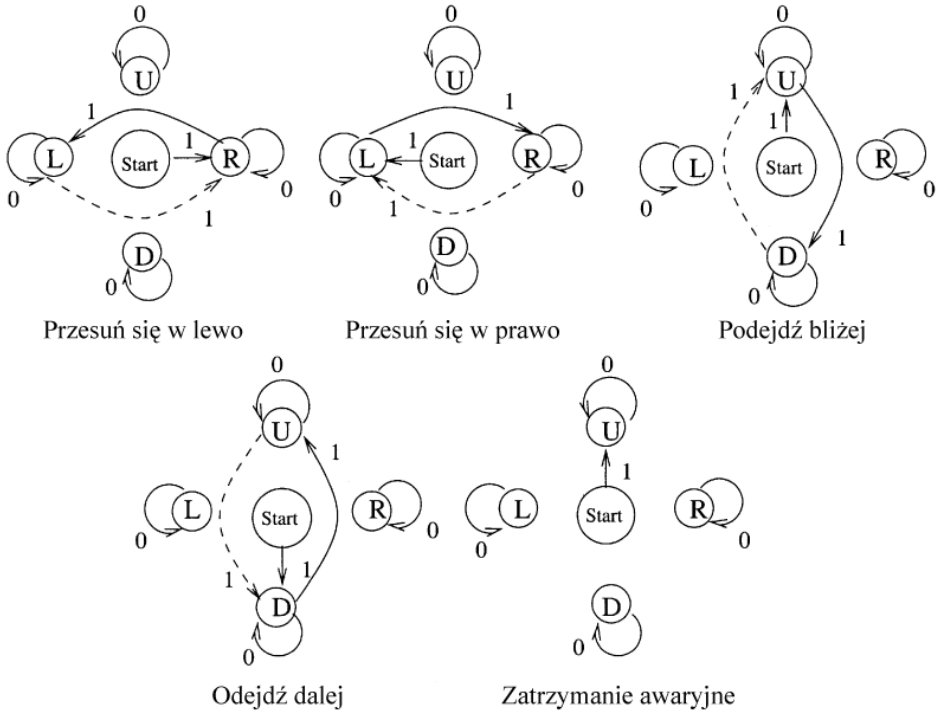


Fig. 3.9 Przykładowe automaty skończone dla pięciu gestów [151]

W systemie opracowanym w ramach rozprawy problem niejednoznaczności segmentacji rozwiązano wykorzystując zamiast metod modelowania stanów mechanizmy logiki rozmytej [153]. Metoda została przedstawiona w rozdziale 5.8.3.

3.2.2 Modelowanie ruchu

Istotnym aspektem modelowania procesów dynamicznych w rozpoznawaniu gestów, poza samym opisem w postaci stanów, jest wnioskowanie na podstawie trajektorii ruchu. Przydatną informację, na podstawie której możliwe jest podejmowanie decyzji o przynależności ruchu do określonej klasy gestu, niesie ze sobą jego prędkość chwilowa, zmiana prędkości i kierunek. Wielkości te wyznaczane są w procesie śledzenia zmian położenia w czasie punktów charakterystycznych, rozmieszczonych na rękach widocznych w strumieniu wizyjnym. W oparciu o punkty dla kolejnych chwil czasu tworzone są lokalne wektory ruchu. W ten sposób cała trajektoria może być wyrażona poprzez sekwencję wektorów [56]. Dla rozpatrywanych w kartezjańskim układzie współrzędnych pozycji (x, y) i kolejnych chwil czasu t i $t - 1$, prędkość υ_t i zmiana

prędkości $\Delta\upsilon_t$ mogą być w prosty sposób wyznaczone zgodnie z zależnościami 3.1 i 3.2.

$$\upsilon_t = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2} \tag{3.1}$$

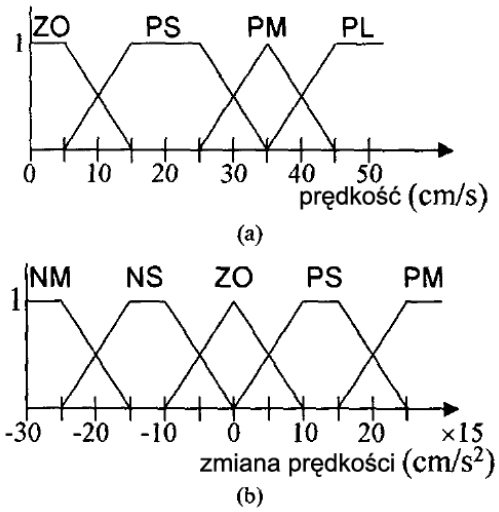
$$\Delta\upsilon_t = \upsilon_t - \upsilon_{t-1} \tag{3.2}$$

Biorąc pod uwagę, że powyższe wielkości w ujęciu modelowanej rzeczywistości przyjmują wartości z dziedziny ciągłej, a modelowane procesy są stochastyczne, do ich reprezentacji warto wykorzystać zbiory rozmyte [153]. Jung-Bae i in. [67] za pomocą zbiorów rozmytych i mechanizmów logiki rozmytej rozpoznają gesty dynamiczne wchodzące w skład koreańskiego języka migowego. Wyróżniają oni pięć faz ruchu, oznaczonych jako: *zatrzymanie*, *przygotowanie*, *atak*, *ruch* i *zakończenie*. Każda z tych faz jest reprezentowana przez prędkość i zmianę prędkości, zdefiniowane różnymi zbiorami rozmytymi, zgodnie z tabelą 3.1. Autorzy wyróżnili sześć zbiorów rozmytych, oznaczonych symbolami ZO, PS, PM, PL, NS i NM, reprezentujących odpowiednio przedziały wartości: *zero*, *dodatnie mała*, *dodatnie średnie*, *dodatnie duże*, *ujemne małe*, *ujemne średnie*.

Tabela 3.1 Reprezentacja poszczególnych faz ruchu za pomocą zbiorów rozmytych definiujących prędkość i zmianę prędkości [67]

Faza ruchu	Warunek	
	Prędkość	Zmiana prędkości
zatrzymanie	ZO	ZO, PS, NS
przygotowanie	PS	PS, NS
atak	PL	PM
ruch	PM	ZO, PS, NS
zakończenie	PS	NM

Zbiory rozmyte przyjęły kształt funkcji trójkątnych i trapezoidalnych zgodnie z rys. 3.10.



Rys. 3.10 Funkcje przynależności dla zbiorów rozmytych reprezentujących: prędkość (a) i zmianę prędkości (b) [67]

Aby dokonać wyboru właściwych metod rozpoznawania gestów dla systemu opracowanego w ramach rozprawy, w pierwszej kolejności celowe było zidentyfikowanie sposobów realizacji systemów przygotowanych przez innych autorów wraz z określeniem ich możliwości, a także problemów, na które napotkali autorzy danego rozwiązania. Z tego względu w rozdziale czwartym dokonano przeglądu wybranych systemów umożliwiających modyfikowanie za pomocą gestów rąk parametrów kształtujących dźwięk.

4 Wybrane metody i systemy kształtowania dźwięku za pomocą gestów

Przegląd literatury w kontekście systemów takich, jak przedstawiony w niniejszej rozprawie, pozwala stwierdzić, że nie są znane rozwiązania umożliwiające bezkontaktową obsługę za pomocą gestów rąk wszystkich kluczowych operacji towarzyszących procesowi miksowania dźwięku. System opracowany w ramach rozprawy ma zatem charakter nowatorski. Znane są jednak systemy, które po rozbudowaniu funkcjonalności mogłyby być wykorzystywane do obsługi procesów miksowania. Marshall i in. w 2009 roku zaprezentowali pracę [92], w której dokonali przeglądu istniejących systemów pozwalających na kontrolowanie za pomocą gestów rąk jednego z ważniejszych elementów każdego zgrania, jakim jest panorama dźwięku. Większość z przedstawionych systemów umożliwia dodatkowo kontrolowanie parametrów związanych z pogłosowością wirtualnego pomieszczenia, w jakim osadzane są panoramowane źródła dźwięku. Systemy te były jednak tworzone z myślą o muzykach, których gesty wykonywane naturalnie w trakcie grania na instrumencie, mogłyby dodatkowo wyzwalać różne funkcje przetwarzające rejestrowany dźwięk i w ten sposób wzbogacać wykonanie. Taka koncepcja oraz urządzenie i system je realizujące zostały przedstawione wcześniej w pracach Modlera [97] [98]. W proponowanym przez Modlera rozwiązaniu użytkownik może sterować parametrami przetwarzania dźwięku poprzez ruch ręki ubranej w rękawicę z czujnikami. Dźwięk przetwarzany jest zgodnie z trzema technikami syntezy: modulacją częstotliwości, fizycznym modelowaniem uderzenia talerza zestawu perkusyjnego oraz syntezą granularną. Śledzone są zmiany 24 kątów pomiędzy kośćmi palców i przyśpieszenie w trzech osiach, co umożliwia rozpoznawanie zarówno gestów statycznych, jak i dynamicznych. Przykładowo, wyciągnięcie palca wskazującego emuluje zdarzenie wciśnięcia przycisku myszy, dzięki czemu taki gest może być użyty do wyboru wirtualnego obiektu skojarzonego z dźwiękiem. Ruch dłoni uformowanej w określony kształt kontroluje w sposób ciągły wartości skojarzonego z tym kształtem parametru. Modyfikacja wartości parametrów zrealizowana jest w oparciu o protokół MIDI. Sposób interakcji z systemem w oparciu o tak zaprojektowany słownik danych można ocenić jako intuicyjny i ergonomiczny. O ergonomii świadczyć może w tym przypadku uniwersalność gestu wyciągnięcia palca wskazującego przyczyniająca się do

eliminacji konieczności zapamiętywania złożonego słownika gestów. Z tego względu, podobny model interakcji przyjęto w trakcie opracowywania systemu przedstawionego w ramach niniejszej rozprawy.

Istotna część prac Modlera dotyczy również warstwy prezentacji informacji. Do wizualizowania interakcji autor wykorzystał język VRML (ang. *Virtual Reality Modelling Language*) oraz interfejs VRML – JAVA. Wizualizowane są zarówno zmiany parametrów przetwarzania dźwięku jak i sam model dłoni, tworzony w sposób dynamiczny na podstawie danych pozyskiwanych z czujników rękawiczki. Poszczególnym parametrom odpowiadają obiekty, nazywane przez autora wirtualnymi obiektami muzycznymi, których cechy charakterystyczne określone są poprzez kolor, wielkość, kształt i pozycję w przestrzeni pseudo-trójwymiarowej. Poza widokiem w postaci światów VRML, w aplikacji dostępny jest również widok, w którym wartości udostępniane przez poszczególne czujniki skorelowane z wartościami kontrolerów MIDI prezentowane są na suwakach. Podobne podejście zastosowano projektując system przedstawiony w niniejszej rozprawie, tj. źródła sygnału mogą być reprezentowane przez obiekty o różnej wielkości i pozycji na ekranie. Dodatkowo wszystkie parametry mogą być wizualizowane w postaci suwaków z wartościami wyświetlanymi u spodu. W systemie Modlera do interpretacji parametrów pozyskiwanych z czujników zastosowano trójwarstwową sztuczną sieć neuronową. Dla podzbioru (nieokreślonego w publikacji autora) słownika gestów średnia skuteczność rozpoznawania wyniosła około 90%.

Należy zauważyć, że wybrana przez Modlera metoda rozpoznawania gestów oparta na zastosowaniu cyber-rękawiczki, pomimo dostarczania precyzyjnych danych opisujących kształt dłoni z uwzględnieniem wielu stopni swobody, według samego autora nie jest pozbawiona wad. Poza oczywistym ograniczeniem swobody ruchów użytkownika wynikającym z istnienia przewodów łączących rękawiczkę z komputerem, głównym problemem związanym z wyborem takiego interfejsu okazał się brak intuicyjności jednoczesnej obsługi kilku parametrów. Mimo technicznych możliwości realizacji takiej funkcjonalności, w praktyce użytkownicy nie byli w stanie z niej skorzystać. Można przypuszczać, że w celu zapewnienia wygodnej obsługi więcej niż jednego parametru jednocześnie, należałoby zastosować rękawiczki na obie dłonie. Powodowałoby to jednak jeszcze większe ograniczenie swobody ruchów użytkownika oraz wzrost kosztów produkcji systemu. Z tych względów na etapie określania założeń projekto-

wych systemu przedstawionego w niniejszej rozprawie wybrano technologię rozpoznawania gestów wykorzystującą przetwarzanie obrazu wizyjnego i dodatkowo założono brak jakichkolwiek czujników. W czasie, kiedy Modler tworzył swój system ówczesny stan techniki nie pozwalał na zastosowanie z powodzeniem metody opartej wyłącznie na analizie obrazu wizyjnego. Zastosowanie manipulatora w postaci rękawiczki było wówczas naturalnym wyborem. Uzasadnieniem rezygnacji ze stosowania tego typu interfejsów obecnie mogą być dodatkowo późniejsze prace Modlera [99] [100], w których interfejs rękawiczki został zastąpiony systemem rozpoznawania gestów w obrazie wizyjnym.

Jako przykład rozwiązania opartego całkowicie na przetwarzaniu obrazu wizyjnego, umożliwiającego rozpoznawanie prostych gestów dłoni i stóp, podać można system Music Maker [45]. System umożliwia przetwarzanie sygnałów muzycznych za pomocą gestów wykonywanych w ramach rehabilitacji ruchowej. Użytkownik poprzez ruch dłoni lub stopy kontroluje takie cechy sygnału, jak poziom i wysokość dźwięku. Możliwe jest wybieranie wyświetlanych na ekranie bloków reprezentujących nuty i tworzenie w ten sposób prostych melodii. Wykonywanym gestom towarzyszy skorelowana z nimi informacja wizualna. Prędkości ruchu odpowiada kolor kwadratu wyświetlanego w środku ciężkości konturu dłoni. Dla nieruchomej dłoni kwadrat przyjmuje kolor czerwony. W miarę wzrostu prędkości, kwadrat zmienia kolor przez żółty po zielony, ten ostatni odnoszący się do szybkiego ruchu. W ten sposób osoba poddana rehabilitacji może odnieść uzyskiwane przez siebie tempo ruchów do tempa docelowego, określonego przez terapeutę. Powiązanie muzyki z taką formą wizualizacji aktywności, według słów autorów, motywuje pacjentów do zwiększania tempa i szybszego osiągania postępów w rehabilitacji w porównaniu do wykonywania ćwiczeń w tradycyjny sposób, tj. bez wsparcia systemem interaktywnym. Autorzy systemu położyli nacisk nie tylko na sposób wizualizowania gestów, ale też na aspekty generowania dźwiękowego sprzężenia zwrotnego z wiernym modelowaniem naturalnej dynamiki gry na rzeczywistym instrumencie. Używając obu dłoni, użytkownik może wyzwalać jednocześnie dźwięki dwóch instrumentów lub tworzyć melodie zawierające dwudźwięki. Powyżej opisana funkcjonalność systemu została osiągnięta przy zastosowaniu prostej metody detekcji obiektów w obrazie wizyjnym, bazującej na modelu koloru. Jeśli obiektem sterującym jest dłoń, system operuje na histogramie koloru skóry. W przypadku obsługi stopą

przed użyciem systemu tworzony jest indywidualny model koloru dla skarpet lub buta. Informacja o ruchu reprezentowana jest przez prędkość i kierunek, wyznaczone na podstawie analizy przemieszczenia punktów charakterystycznych określających pozycje dłoni w kolejnych ramkach obrazu. Takie samo podejście do modelowania ruchu przyjęto w niniejszej rozprawie. W szczególności, wektory prędkości i kierunku stanowią dobrą podstawę do modelowania ruchu w języku logiki rozmytej.

Spośród interfejsów umożliwiających przetwarzanie dźwięków muzycznych za pomocą gestów rąk wymienić można również system WAVE (*Virtual Audio Environment*) autorstwa Valboma i in. [140]. Rozwiązanie to umożliwia wyzwalanie za pomocą gestów rąk wybranych pętli muzycznych lub odtwarzanie dźwięków skal chromatycznych. Ruch rąk odzwierciedla ruch wirtualnych różdżek na ekranie, za pośrednictwem których możliwe jest wybieranie i zmiana położenia obiektów. System kontrolowany jest za pośrednictwem myszy wyposażonej w trójosiowe czujniki. Poprzez zastosowanie technologii wirtualnej rzeczywistości i systemu dźwięku przestrzennego opartego na dwukanałowym odsłuchu bliskiego pola zintegrowanym z systemem odsłuchowym 4.1 osiągnięto zanurzenie użytkownika w przestrzeni trójwymiarowej.

Możliwości, jakie oferuje technologia wyświetlania obrazu trójwymiarowego zostały wykorzystane w kontekście przetwarzania dźwięków muzycznych za pomocą gestów również przez autorów innej pracy [12]. Berthaut i in. bazując na pracach Cadoza [21] dotyczących gestów muzycznych wyszczególnili trzy grupy gestów, tj. gesty selekcji – służące do wyboru obiektów reprezentujących dźwięki, gesty modulacji – odpowiedzialne za przetwarzanie dźwięków i gesty pobudzenia – służące do wyzwalania dźwięków [12]. Wybieranie i przesuwanie obiektów graficznych osadzonych w przestrzeni trójwymiarowej reprezentujących zdarzenia dźwiękowe, realizowane jest za pomocą wirtualnych promieni świetlnych prowadzonych od rąk w kierunku obiektów. Taki sposób interakcji znajduje zastosowanie w wielu rozwiązaniach z dziedziny wirtualnej rzeczywistości [16]. Autorzy dużą wagę przywiązali do zapewnienia możliwości jednoczesnej kontroli wielu parametrów. W tym celu oparli system na koncepcji interakcji za pomocą reaktywnych narzędzi (ang. *reactive widgets*) [87]. Poprzez zastosowanie techniki promieni świetlnych, możliwe jest jednoczesne wybranie za pomocą obu dłoni dwóch narzędzi reaktywnych. Modyfikacji dźwięku dokonuje się poprzez zmianę wyglądu elementu graficznego, wprowadzając go w tzw. tunel, tym samym wpływając na jego

kształt. Autorzy zbadali dwie techniki wyzwalania dźwięków, tj. technikę polegającą na zastosowaniu popularnego kontrolera Wii [104] i technikę opartą na doprowadzaniu do kolizji wirtualnych obiektów na ekranie. Kontroler Wii wyposażony jest w akcelerometry mierzące przyśpieszenie w osi poziomej (lewo – prawo) i pionowej oraz obrót w płaszczyźnie horyzontalnej, czujnik światła podczerwonego, osiem przycisków dwustanowych oraz kontroler krzyżakowy (ang. *Directional pad – D-pad*). Autorzy pracy uznali kontroler za ograniczający ekspresję i niewystarczająco precyzyjny w zachowaniu zgodności akcji z tempem przetwarzanej za pomocą gestów muzyki. Wyzwalanie dźwięków wskutek zderzania ze sobą obiektów uznane zostało za bardziej ekspresyjne, ponieważ prędkość zderzania może być jednym z czynników wpływających na przetwarzanie dźwięku. Ta technika jednak również okazała się niesatysfakcjonująca pod względem precyzji czasowej. Autorzy dokonali przeglądu kontrolerów sprzętowych nadających się potencjalnie do obsługi procesów modyfikowania dźwięków muzycznych i uznali za konieczne opracowanie nowego rodzaju sprzętowego interfejsu. Pomimo, że przedstawiona praca powstała w dobie rozwoju technologii wizyjnych i mniejszego zainteresowania interfejsami sprzętowymi, oparcie interakcji w całości na systemie wizyjnym nie było przez autorów brane pod uwagę. Warto zauważyć, że autorzy wyrażają się krytycznie na temat opisanego wcześniej systemu WAVE. Możliwości w zakresie wizyjnego sprzężenia zwrotnego i interakcji oceniali jako ograniczone, co ich zdaniem wynikało ze specyfiki emulowania kontrolerów sprzętowych. Zdaniem autorów rozwiązanie to oferuje nowe techniki interakcji i prezentuje interesujące podejście do wizualizowania informacji, ale nie wykorzystuje jednak w pełni możliwości środowisk zanurzenia w rzeczywistości wirtualnej. Uwaga ta dotyczy również innych systemów oferujących nowatorski sposób interakcji i wizualizacji w procesie kształtowania dźwięku [51] [93] [103] [105] [106]. Istotna uwaga autorów, z punktu widzenia tematyki niniejszej rozprawy, dotyczy faktu, że żaden z tych systemów nie oferuje możliwości jednoczesnej kontroli wielu parametrów.

Przydatność gestów w obsłudze systemów DAW zauważyli Balin i Loviscach [7]. Zaproponowali oni sterowanie wspomagane gestami wykonywanymi za pomocą myszy. Rozwiązanie to wzorowane jest na funkcjonalności udostępnianej przez popularne przeglądarki internetowe [30]. Wykonanie myszą ruchu zdefiniowanego w słowniku gestów powoduje wywołanie przypisanej funkcji. Większość operacji dotyczy procesu

edycji materiału dźwiękowego oraz funkcji widoku. Spośród funkcji związanych z edycją wymienić można: dzielenie regionu, wycinanie i kopiowanie, klejenie regionów czy duplikowanie. Do funkcji widoku należą między innymi: dopasowywanie zaznaczonego regionu lub ścieżki do rozmiaru okna, przywoływanie ustawień powiększania. Autorzy nie przewidzieli możliwości wywoływania za pomocą gestów operacji miksowania sygnałów fonicznych. Ograniczyli się jedynie do zdefiniowania dwóch gestów oznaczających zmniejszenie i zwiększenie wartości parametru. Dla tych dwóch gestów, w przeciwieństwie do pozostałych, akcja nie ogranicza się jedynie do jednorazowego wywołania funkcji, ale do ciągłej kontroli parametru. Warto zwrócić uwagę na fakt, że w ramach przeprowadzonego przeglądu badań w dziedzinie automatycznego rozpoznawania gestów, autorzy podkreślają zwiększenie znaczenia wizualnej informacji zwrotnej w przypadku, gdy interakcji nie towarzyszą zwrotne bodźce czuciowe (np. zmiana pozycji, w jakiej znajduje się przycisk). Wnioski te oparli m.in. na pracach Buxtona i in. [20].

W kontekście niniejszej rozprawy, zauważyć można, że zastosowanie gestów wykonywanych myszą do wywoływania i obsługi operacji miksowania dźwięku pozwoliłoby na wykluczenie zmysłu wzroku z procesu percepcji. Poprzez odpowiednio przygotowane testy możliwe by było w oparciu o taki system sprawdzenie wpływu obecności informacji wizualnych na decyzje podejmowane w procesie miksowania. System obsługiwany jedynie za pomocą myszy nie oferowałby jednak możliwości jednoczesnej edycji więcej niż jednego parametru. Z tego względu model sterowania oparto na interfejsie sterowania komputerem za pomocą gestów zaprojektowanym w kontekście specyfiki miksowania dźwięku. Dzięki temu możliwe stało się porównanie ergonomii interfejsu sterowania za pomocą gestów z ergonomią pracy z wykorzystaniem myszy komputerowej w procesie miksowania dźwięku i sprawdzenie jej wpływu na ocenę subiektywną zgrań w porównaniu z wpływem obecności lub braku informacji wizualnej reprezentującej wartości parametrów.

System, który umożliwia miksowanie sygnałów fonicznych w sposób zdalny za pomocą ruchów wykonywanych w powietrzu zaprezentowali Selfridge i Reiss [126]. Wykorzystano w nim kontroler Wii [104]. Za pomocą ruchów kontrolera i przycisków możliwa jest modyfikacja parametrów cyfrowego przetwarzania sygnału. Autorzy systemu przeprowadzili badania nad możliwością zastosowania czujników podczerwieni, w które wyposażony jest kontroler do rozszerzenia możliwości zdalnej obsługi proce-

sów miksowania. Pomimo popularności rozwiązań realizujących rozpoznawanie gestów w oparciu o kamery i diody podczerwieni, stwierdzili oni, że w obszarze miksowania dźwięku tego typu podejście nie znajduje praktycznego zastosowania. Powodem takiego stanu rzeczy jest ograniczenie przestrzeni ruchów wykonywanych kontrolerem zapewniające nieprzerwany odbiór wiązki światła podczerwonego przez czujnik. Dodatkowo wykorzystanie światła podczerwonego przeczyło założeniu o możliwości zajmowania przez użytkownika dowolnej pozycji odsłuchowej. Znaczący problem, zauważony przez autorów systemu w trakcie jego testowania, związany jest z zakresem czułości akcelerometrów. Wykonanie subtelnego ruchu nie powodowało zarejestrowania zmiany przez czujniki, co sprawia, że precyzyjna modyfikacja wartości parametrów jest utrudniona. W kontekście niniejszej rozprawy spostrzeżenia te uzasadniają opracowany model sterowania, w którym nie wykorzystuje się diod i czujników światła podczerwonego ani akcelerometrów.

Na podstawie dokonanego powyżej przeglądu systemów kształtowania dźwięku za pomocą gestów, jak i przeglądu metod rozpoznawania gestów zawartego w rozdziale 3., sformułowano założenia projektowe systemu, będącego tematem rozprawy. Założenia te przedstawiono w kolejnym rozdziale.

5 System miksowania dźwięku za pomocą gestów rąk

5.1 Założenia projektowe

Założenia projektowe systemu określono, mając na względzie przede wszystkim trzy aspekty. Po pierwsze, przyjęto, że w oparciu o wytworzony system musi być możliwe sprawdzenie czy odzwierciedlanie zmian parametrów fonicznych w postaci wizualnej wpływa w istotny sposób na miksowanie sygnałów. Po drugie, założono, że interfejs systemu powinien być zaprojektowany w taki sposób, aby obsługa za pomocą gestów była łatwa i intuicyjna, tym samym ergonomiczna i skuteczna. W ramach tego założenia przyjęto, że system powinien wspierać rozpoznawanie gestów obu rąk, umożliwiając tym samym jednoczesną edycję dwóch parametrów. Wykorzystane do obsługi gesty statyczne i dynamiczne powinny być łatwe do wykonania. Funkcje systemu związane ze złożonymi gestami dynamicznymi powinny być dostępne również w postaci ikonograficznej. W powszechnych sposobach miksowania sygnałów fonicznych, w których wykorzystuje się stół mikserski bądź kontroler, występuje problem „zakolorowań" dźwięku powstających wskutek odbić od powierzchni płaskich (blat stołu lub konsoleta). Brak konieczności stosowania takich urządzeń w interakcji za pomocą gestów pozwolił na sformułowanie trzeciego założenia podstawowego, zgodnie z którym system powinien zapewniać warunki odsłuchowe, w których dźwięk pozbawiony jest „zakolorowań". Założono też, że w ramach eksperymentów wstępnych autor rozprawy powinien dokonać wyboru optymalnego algorytmu klasyfikującego gesty statyczne oraz na podstawie uzyskanej skuteczności wybranej metody odpowiedzieć na pytanie czy jest możliwa obsługa systemu bez przeprowadzania procesów kalibracji.

W związku z pierwszym założeniem wszystkie operacje związane z przetwarzaniem dźwięku mogą być wykonywane z wyłączeniem zwrotnych informacji wizualnych, chociaż opcja pracy z pełnym interfejsem graficznym jest również dostępna. Dodatkowo, niezależnie od pracy w trybie ograniczonego lub pełnego interfejsu graficznego, dostępny jest widok pasków menu z ikonograficznymi przyciskami reprezentującymi operacje miksowania dźwięku. W ten sposób, w zależności od potrzeb użytkownika, możliwe jest wywoływanie funkcji systemu za pomocą gestów lub wybieranie ich poprzez skierowanie ręki nad odpowiednią ikonę.

Zamierzeniem autora rozprawy było stworzenie systemu, który, poza wyznaczeniem ergonomicznego sposobu obsługi procesów miksowania za pomocą gestów, zapewniałby warunki umożliwiające wiarygodne sprawdzenie wpływu bodźców wzrokowych na decyzje podejmowane w trakcie kreowania podstawowych elementów każdego miksu, tj. balansu (poziomy i relacje częstotliwościowe), szerokości bazy, dynamiki i głębi. Powyższe pojęcia, ogólnie znane w inżynierii dźwięku, wyjaśnione zostały w Słowniku pojęć. Jednocześnie narzędzie mogłoby stanowić podstawy nowego, alternatywnego trendu w dziedzinie projektowania systemów produkcji nagrań muzycznych, ukierunkowanych w większym stopniu na aspekty percepcyjne. Założono również, że z punktu widzenia inżynierów dźwięku istotne jest porównanie ergonomii miksowania za pomocą gestów z ergonomią miksowania przy użyciu myszy i klawiatury. Nie było natomiast zamierzeniem autora stworzenie systemu, który mógłby konkurować z istniejącymi na rynku systemami DAW pod względem liczby i jakości wykorzystywanych algorytmów przetwarzania sygnałów. Przyjęto zatem założenie, że system powinien funkcjonować jako sterownik wybranej aplikacji DAW. Tym samym wszelkie operacje przetwarzania sygnału fizycznie wykonywane są w środowisku aplikacji DAW, natomiast ich rezultat jest widoczny w interfejsie opracowanym w ramach rozprawy. Dodatkowo przyjęto założenie o ograniczeniu zbioru edytowanych parametrów do następujących: poziomu, panoramy, wzmocnienia górno-zakresowego korektora częstotliwości, progu i stopnia kompresora dynamiki oraz czasu pogłosu i stosunku dźwięku pogłosowego do bezpośredniego. Parametry te są bezpośrednio związane z elementami miksu wymienionymi wcześniej.

5.1.1 Wymagania stawiane systemowi

Wymagania funkcjonalne i pozafunkcjonalne dotyczące systemu zostały zebrane w formie dokumentu specyfikacji i analizy wymagań umieszczonego w dodatkach do rozprawy w sekcji G. Poniżej wyszczególniono najistotniejsze wymagania z punktu widzenia wytworzenia systemu w taki sposób, aby możliwe było wiarygodne sprawdzenie wpływu informacji wizualnej i ergonomii obsługi na edycję wartości parametrów.

Wymagania dotyczące funkcjonalności

Przyjęto, że w zakresie wykonywania funkcji związanych bezpośrednio z miksowaniem materiału muzycznego interfejs użytkownika powinien umożliwiać alternatywnie sterowanie za pomocą ruchów i gestów rąk (palców dłoni, przedramion), albo za pomocą klawiatury i myszy. Funkcjonalność ta jest konieczna w kontekście przeprowadzenia badań mających na celu porównanie wyników procesu miksowania za pomocą komputera z rezultatami otrzymanymi za pomocą opracowanego interfejsu. Dopuszczono możliwość, aby funkcje niezwiązane bezpośrednio z procesami miksowania, takie jak np. załadowanie sygnałów fonicznych do systemu lub eksport sumy sygnałów do pliku dźwiękowego były wykonywane jedynie za pomocą klawiatury i myszy.

Interfejs został zaprojektowany w taki sposób, aby umożliwiał sterowanie za pomocą intuicyjnych i łatwych do zapamiętania gestów. Przyjęto założenie o ograniczeniu słownika gestów do niezbędnego minimum. Wymaganie to podyktowane było koniecznością zapewnienia łatwości nauki obsługi interfejsu oraz wysokiej skuteczności klasyfikacji gestów poprzez dużą separowalność pomiędzy klasami. Jednocześnie uznano, że zbiór gestów powinien być na tyle złożony, aby obsługa interfejsu nie sprowadzała się jedynie do wyboru i edycji parametrów poprzez kaskadowe menu obsługiwane przez umieszczenie dłoni nad odpowiednią ikoną. Przyjęto założenie, że kąt obrotu dłoni przy gestach statycznych nie powinien być czynnikiem decydującym o przynależności kształtu do określonej klasy gestów. Założenie to było podyktowane zamiarem zapewnienia dużej wygody użytkowania systemu.

Przyjęto, że system powinien umożliwiać użytkownikowi przeprowadzenie wszystkich procesów miksowania w dwóch trybach interfejsu graficznego, tj. w trybie pełnym i ograniczonym. W trybie pełnym wszelkie operacje są odzwierciedlane za pomocą zmian wyglądu elementów graficznych. W trybie ograniczonym wyświetlane są jedynie niezbędne informacje. Informacje te odpowiadają wszelkim operacjom, które nie posiadają odzwierciedlenia w postaci bodźca dźwiękowego. Dodatkowo udostępniana jest informacja wizualna dla operacji, o których wykonaniu użytkownik mógłby zapomnieć w trakcie użytkowania systemu. Operacje te dotyczą funkcji transportu (odtwarzanie, przewijanie, zatrzymanie) oraz wyciszania ścieżki i ustawiania jej w tryb sa-

modzielnego odtwarzania (*solo*). Dodatkowo interfejs sygnalizuje aktualnie wybrane sygnały foniczne i parametry dla lewej i prawej ręki.

Wymagania dotyczące wyglądu elementów interfejsu

Ponieważ metoda detekcji rąk bazuje na odejmowaniu obrazu pozyskanego z kamery od obrazu wyświetlanego przez rzutnik, istotne jest zachowanie dużego kontrastu pomiędzy cieniem rąk a grafiką interfejsu. Stopień kontrastu warunkuje skuteczność detekcji gestów a w konsekwencji skuteczność ich rozpoznawania. Z tego względu, jako kolor tła elementów interfejsu graficznego wybrano biel.

Rozmiar wszystkich elementów graficznych, takich jak przyciski czy piktogramy, służących do wywoływania operacji systemu, został dobrany w taki sposób, aby możliwa była bezproblemowa obsługa za pomocą rąk. Oznacza to, że system zapewnia kompensowanie ewentualnych niedokładności w odzwierciedleniu położenia ręki w pozycji kursora na ekranie. Sposób użycia interfejsu opisano w rozdziałach 5.4 i 5.5.

5.2 Wybór metody klasyfikacji gestów rąk

Wiarygodne rozpoznawanie gestów w obrazie wizyjnym wymaga typowo użycia algorytmów przetwarzania obrazu i klasyfikacji zdarzeń o dużej złożoności czasowej [3] [96] [142]. Zastosowanie metody przetwarzania obrazu złożonej z wielu algorytmów pozwala otrzymać obraz zawierający jednoznaczną informację o lokalizacji ręki i jej kształcie. W połączeniu z kaskadą algorytmów klasyfikacji obiektów i zdarzeń teoretycznie osiągnąć można wysoką skuteczność rozpoznawania gestów. Zastosowanie wielu algorytmów o dużej złożoności czasowej powoduje jednak spadek wydajności całego systemu w przypadku zastosowania powszechnie dostępnych komputerów. Z tego względu w niektórych zastosowaniach użycie takiej technologii jest niepraktyczne [96] [142]. System bowiem nie jest w stanie zapewnić wystarczającej prędkości przetwarzania klatek strumienia wizyjnego. W efekcie szybkie ruchy towarzyszące gestom dynamicznym lub szybkie zmiany kształtów związane z gestami statycznymi nie są w pełni rejestrowane. Efektem może być błędna interpretacja wykonanego gestu. Dodatkowo problemem jest użycie takiego systemu w interakcji, w której wykonywanie gestu wpływa na stan systemu w sposób ciągły, a szybkość odzwierciedlenia ruchu w postaci zmiany stanu ma szczególne znaczenie. Przykładem zastosowania, w którym występują

oba poruszone aspekty jest rozpatrywane w ramach niniejszej rozprawy miksowanie dźwięku. Z tego względu wybór metod w oparciu, o które został zbudowany system jest podyktowany poszukiwaniem kompromisu pomiędzy wiarygodnością (skutecznością) detekcji rąk w obrazie i klasyfikacją gestów a wydajnością systemu. Kolejnym, oczywistym czynnikiem decydującym o wyborze określonych metod i odrzuceniu innych była konieczność spełnienia przedstawionych w niniejszym rozdziale założeń projektowych.

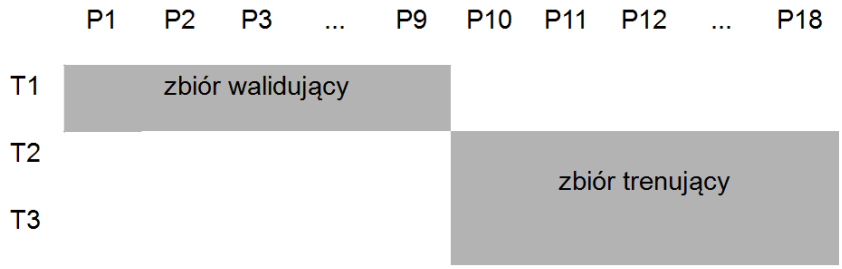
5.2.1 Badanie skuteczności rozpoznawania gestów statycznych

Skuteczność rozpoznawania gestów statycznych w komputerowych systemach wizyjnych w dużej mierze zależy od zastosowanej metody klasyfikacji kształtów. Wybór metody determinuje globalną skuteczność detekcji gestów, łatwość nauki obsługi systemu i jego użyteczność. Z tego względu istotne jest, aby wybrać optymalną metodę dla danego problemu. W tym celu w ramach wstępnych eksperymentów zbadano następujące klasyfikatory: algorytm z grupy algorytmów najbliższych sąsiadów z niezagnieżdżoną generalizacją (NNge), naiwną sieć Bayesa, drzewo C4.5 (J48), drzewo losowe, lasy losowe, sztuczne sieci neuronowe i maszyny wektorów nośnych (SVM). Klasyfikatory te zostały wybrane ze względu na wysoką skuteczność, osiągniętą dla podobnych problemów, wykazaną w badaniach opisanych w literaturze [11] [14] [18] [46] [74] [110] [117] [147]. Implementacja klasyfikatorów pochodziła z systemu WEKA [57]. Powodem wyboru tego systemu była możliwość stworzenia w oparciu o jego klasy własnej aplikacji testowania klasyfikatorów. W ten sposób automatycznie zbadano różne metody dla różnych wektorów parametrów. Ze względu na złożoność zagadnień poruszonych w ramach rozprawy w niniejszym rozdziale przywołano jedynie kluczowe parametry wyżej wymienionych klasyfikatorów bez przedstawiania teorii leżących u ich podstaw.

Skuteczność każdego z klasyfikatorów została sprawdzona w oparciu o nagrane sekwencje gestów 18 osób (5 kobiet i trzynastu mężczyzn). Osoby wykonywały 4 gesty statyczne, każdy dla trzech różnych cyklicznych trajektorii ruchu, tj. dla przemieszczania ręki lewo – prawo, góra – dół oraz dla zakreślania okręgu. Dla każdej trajektorii pozyskiwano 30 klatek reprezentujących gest statyczny. Reprezentację pojedynczego gestu statycznego zawierało zatem 90 klatek (3 trajektorie po 30 klatek). Kształt dłoni

parametryzowany był za pomocą histogramu PGH (ang. *Pairwise Geometrical Histograms*) [17]. Pojedynczy histogram składał się z 15 przedziałów klasowych o maksymalnej liczebności równej 10.

Trzy zbiory histogramów reprezentujących kształty uzyskane dla trzech trajektorii ruchu posłużyły do utworzenia zbiorów trenujących i testujących z podziałem 2/3. Dla testów wstępnych, mających na celu wyłonienie najlepszych klasyfikatorów w krótkim czasie, zastosowano metodę dwukrotnej walidacji krzyżowej. Zgodnie z rys. 5.1 połowa zbiorów zawierających histogramy dla pierwszej trajektorii ruchu utworzyła zbiór walidujący, natomiast połowa zbiorów zawierających histogramy dla pozostałych dwóch trajektorii ruchu utworzyła zbiór trenujący.



Rys. 5.1 Sposób podziału zbiorów na trenujące i walidujące w metodzie dwukrotnej walidacji krzyżowej (TX – identyfikator trajektorii, PX – identyfikator osoby)

Przyjęta metoda podziału zbiorów na trenujące i walidujące pozwoliła uzyskać wysoką dyskryminację pomiędzy skutecznościami klasyfikatorów. Jednocześnie metoda ta charakteryzowała się dużo niższym kosztem obliczeniowym niż klasyczna metoda *k*-krotnej walidacji krzyżowej lub jej odmiana, w której *n*-elementowa próba jest dzielona na *n* podzbiorów, zawierających po jednym elemencie (ang. *leave-one-out*). W celu zbadania skuteczności i wydajności klasyfikatorów autor rozprawy opracował aplikację w języku Java wykorzystującą klasy systemu WEKA.

Dla każdego klasyfikatora, parametry zapewniające najwyższą skuteczność określono wykorzystując metodę *grid-search* [61]. Sekwencje wartości dla pojedynczych parametrów klasyfikatorów zostały wyznaczone zgodnie z funkcją 2^k, $k \in [-m, n]$, gdzie k, m, n to liczby całkowite, wybrane w oparciu o przegląd literaturowy [11] [14] [18] [46] [74] [117] [147] i w drodze zgrubnej weryfikacji empirycznej. Dla klasyfikatora NNge

parametry określające liczbę folderów dla informacji wspólnych *i* oraz liczbę prób generalizacji *g*, przyjęły wartości z przedziału $[2^0, 2^4]$. Naiwna sieć Bayesa została zbadana dla trzech warunków, tj. z wykorzystaniem estymacji jądrowej dla wartości histogramów reprezentujących kształty dłoni, dla rozkładu normalnego wartości histogramów i z wykorzystaniem nadzorowanej dyskretyzacji w celu konwersji numerycznych wartości histogramów do cech nominalnych. Dla algorytmu J48, będącego implementacją klasyfikatora C4.5 w systemie WEKA, współczynnik pewności *C* zastosowany w redukcji drzew przyjmował wartości z przedziału $[2^{-17}, 2^0]$, a minimalna liczba instancji w liściu drzewa *m* zawierała się w przedziale $[2^1, 2^6]$. Dla drzewa losowego liczba losowo wybranych histogramów k_{rt} przyjmowała wartości z przedziału $[2^0, 2^{10}] \cup \{\log_2 n_{rt} + 1\}$, gdzie n_{rt} oznacza liczbę histogramów, a minimalna całkowita waga instancji w liściu *m* zawierała się w przedziale $[2^{-17}, 2^0]$. Dla losowego lasu, liczba drzew przyjmowała wartości z przedziału $[2^0, 2^{10}]$, a liczba losowo wybranych histogramów k_{rf} zawierała się w przedziale $[2^0, 2^5] \cup \{\log_2 n_{rf} + 1\}$, gdzie n_{rf} oznacza liczbę histogramów. Dla sztucznej sieci neuronowej, współczynnik uczenia *l* i moment *m* zawierały się w przedziale $[2^{-4}, 2^0]$, natomiast liczba epok *e* przyjmowała wartości z przedziału $[2^2, 2^5]$. Dla maszyny wektorów nośnych, wartość funkcji kosztu wybrano z przedziału $[2^0, 2^{14}]$, natomiast współczynnik gamma przyjął wartości z przedziału $[2^{-15}, 2^{-5}]$. Wartości tych użyto zarówno dla jądra liniowego, jak i jądra RBF (ang. *Radial Basis Function*).

5.2.2 Wyniki eksperymentów wstępnych

Wyniki badania klasyfikatorów, w kolejności od najlepszego do najsłabszego, osobno dla lewej i prawej dłoni, dla parametrów zapewniających najwyższą skuteczność przedstawiono w tabelach 5.1 i 5.2.

Tabela 5.1 Wyniki badania klasyfikatorów dla lewej dłoni

	Średnia skuteczność [%]	Średni czas treningu [ms]	Średni czas walidacji [ms]	Parametry
SVM (LibSVM, liniowe jądro)	92,82	2599	1123	$\gamma = 2^{-15}$, $C = 2^{1}$
SVM (LibSVM, jądro RBF)	92,82	2508	1159	$\gamma = 2^{-11}$, $C = 2^{11}$
SSN	91,67	1458	187	$l = 2^{-3}$, $m = 2^{-5}$, $e = 2^{3}$, jedna warstwa ukryta, 4 perceptrony w warstwie
Losowy las	89,91	59644	722	$i = 2^{9}$, $k = 2^{4}$, nieograniczona głębokość
NNge	83,47	14234	8073	$g = 2^{2}$, $i = 2^{4}$
Naiwna sieć Bayesa	79,49	303	73	z nadzorowaną dyskretyzacją
C4.5 (J48)	77,73	1342	4	$C = 2^{-7}$, $m = 2$
Losowe drzewo	77,04	443	3	$k = 2^{6}$, $m = 2^{-17}$

Tabela 5.2 Wyniki badania klasyfikatorów dla prawej dłoni

	Średnia skuteczność [%]	Średni czas treningu [ms]	Średni czas walidacji [ms]	Parametry
SSN	91,11	3854	181	$l = 2^{-3}$, $m = 2^{-3}$, $e = 2^{5}$, jedna warstwa ukryta, 4 perceptrony w warstwie
SVM (LibSVM, liniowe jądro)	88,52	2811	1430	$\gamma = 2^{-8}$, $C = 2^{2}$
SVM (LibSVM, jądro RBF)	88,52	5166	2194	$\gamma = 2^{-6}$, $C = 2^{2}$
Losowy las	84,44	15047	112	$i = 2^{7}$, $k = 2^{4}$, nieograniczona głębokość
Naiwna sieć Bayesa	78,56	189	5231	z estymacją jądrową
Losowe drzewo	77,92	993	4	$k = 2^{8}$, $m = 2^{-17}$
C4.5 (J48)	76,44	1292	4	$C = 2^{-6}$, $m = 2^{3}$
NNge	75,42	8610	5409	$g = 2^{3}$, $i = 2^{2}$

Klasyfikatorami, dla których osiągnięto najwyższą skuteczność rozpoznawania gestów były: sztuczna sieć neuronowa i maszyna wektorów nośnych. Dla klasyfikatora SVM, niezależnie od wyboru rodzaju jądra, dla odpowiednich wartości funkcji kosztu i parametru γ, uzyskano identyczną skuteczność. Jest to zgodne z obserwacjami zawartymi w literaturze [61], z których wynika, że istnieją wartości parametrów, dla których jądro RBF zapewnia co najmniej taką skuteczność klasyfikacji jak jądro liniowe.

Różnice skuteczności klasyfikacji kształtów lewej i prawej dłoni potwierdzają zjawisko zaobserwowane i udokumentowane w trakcie innych praz prowadzonych w Katedrze Systemów Multimedialnych [74]. Wyższa skuteczność dla dłoni lewej związana jest prawdopodobnie z tym, że grupa testowa w większości składała się z osób praworęcznych, dla których formowanie danego kształtu przez dłoń lewą jest mniej naturalne niż dla dłoni prawej. Prowadziło to do wykonywania gestów lewej dłoni z większą starannością. W konsekwencji uzyskano lepszą separowalność pomiędzy klasami gestów, co spowodowało wzrost skuteczności klasyfikacji.

Dla klasyfikatora SVM i sztucznej sieci neuronowej dodatkowo sprawdzono skuteczność za pomocą metody walidacji *leave-one-out*. Zbiór walidujący tworzyły histogramy reprezentujące kształty pozyskane przy zadanej trajektorii ruchu ręki pojedynczej osoby. Histogramy dla pozostałych dwóch trajektorii ruchu rąk pozostałych osób utworzyły zbiór trenujący. Taka metoda testowania, w przeciwieństwie do typowej metody walidacji krzyżowej z losowym podziałem danych na trenujące i walidujące, pozwoliła ocenić zdolność generalizowania klasyfikatorów. Wysoka zdolność generalizowania pozwalałaby bowiem na zastosowanie w projektowanej aplikacji systemu klasyfikacji wytrenowanego w oparciu o zbiór nagrań, który posłużył do przeprowadzenia opisywanych testów. Tym samym, użytkownik systemu byłby zwolniony z fazy kalibracji polegającej na trenowaniu klasyfikatorów. Wyniki badań skuteczności metodą *leave-one-out* klasyfikatora SVM i sztucznej sieci neuronowej przedstawiono w tabelach 5.3–5.5.

Tabela 5.3 Wyniki badania klasyfikatora SVM z liniowym jądrem metodą walidacji leave-one-out

	Lewa dłoń	Prawa dłoń
Maksymalna skuteczność [%]	100,00	100,00
Minimalna skuteczność [%]	66,67	65,83
Średnia skuteczność [%]	95,68	94,65
Mediana [%]	98,33	97,50
Odchylenie standardowe	6,43	7,82
Wariancja	40,15	59,34
Dolna granica 95% przedziału ufności	93,85	92,42
Górna granica 95% przedziału ufności	97,51	96,87
Skośność	-2,30	-2,07
Kurtoza	6,14	4,12
Średni czas treningu [ms]	6435	6598
Średni czas walidacji [ms]	197	203

Tabela 5.4 Wyniki badania klasyfikatora SVM z jądrem RBF metodą walidacji *leave-one-out*

	Lewa dłoń	Prawa dłoń
Maksymalna skuteczność [%]	100,00	100,00
Minimalna skuteczność [%]	65,83	65,83
Średnia skuteczność [%]	95,56	94,09
Mediana [%]	98,33	98,33
Odchylenie standardowe	6,59	7,97
Wariancja	42,16	61,75
Dolna granica 95% przedziału ufności	93,68	91,82
Górna granica 95% przedziału ufności	97,43	96,36
Skośność	-2,28	-1,47
Kurtoza	6,03	1,50
Średni czas treningu [ms]	6575	10478
Średni czas walidacji [ms]	204	335

Tabela 5.5 Wyniki badania sztucznej sieci neuronowej metodą walidacji leave-one-out

	Lewa dłoń	Prawa dłoń
Maksymalna skuteczność [%]	100,00	100,00
Minimalna skuteczność [%]	60,83	69,17
Średnia skuteczność [%]	92,53	93,24
Mediana [%]	96,25	95,83
Odchylenie standardowe	7,99	7,36
Wariancja	61,79	52,64
Dolna granica 95% przedziału ufności	90,26	91,15
Górna granica 95% przedziału ufności	94,80	95,33
Skośność	-1,56	-1,16
Kurtoza	2,84	0,68
Średni czas treningu [ms]	2386	2379
Średni czas walidacji [ms]	21	21

Duża wartość mediany i wariancji oraz ujemna skośność wskazują na to, że w skład grupy testowej wchodziła osoba lub kilka osób, dla których wykonywane gesty znacząco różniły się od innych. Dla tych osób skuteczność klasyfikatorów, wytrenowanych na zbiorach zawierających sparametryzowane kształty dłoni pozostałych uczestników testów, została znacząco obniżona. Na podstawie tej obserwacji przyjęto, że zastosowany w systemie klasyfikator powinien być trenowany w oparciu o indywidualne kształty dłoni użytkownika pozyskiwane w fazie kalibracji. Było to jedno z głównych założeń projektowych sformułowanych po przeprowadzeniu wstępnych badań.

Dla klasyfikatora SVM z liniowym jądrem przeprowadzono dodatkowo precyzyjne poszukiwanie parametrów zapewniających najwyższą skuteczność klasyfikacji. Bazując na doniesieniach literaturowych [61], można stwierdzić, że ponowne zastosowanie metody *grid-search* z mniejszym krokiem na zawężonych przedziałach wokół wartości wskazanej w pierwszej fazie jako zapewniającej najwyższą skuteczność, pozwala zidentyfikować parametry, dla których skuteczność klasyfikatora wzrasta o 0,1%. Wartości parametrów, dla których uzyskano taki wzrost wyniosły odpowiednio: $C = 2^{1,25}$, $\gamma = 2^{-15}$ dla klasyfikatora kształtów lewej dłoni i $C = 2^{2}$, $\gamma = 2^{-8,75}$ dla klasyfikatora kształtów prawej dłoni.

5.2.3 Wybór metod

Odnosząc się do uzyskanych wyników w eksperymentach wstępnych, najwyższe skuteczności uzyskano dla maszyn wektorów nośnych i sztucznej sieci neuronowej. Kolejnym w szeregu najlepszych klasyfikatorów okazał się las losowy. Klasyfikator ten odrzucono jednak ze względu na nieakceptowalnie długi czas treningu. Pomimo porównywalnych skuteczności uzyskanych dla maszyn wektorów nośnych i sztucznej sieci neuronowej oraz krótszego czasu walidacji drugiego z klasyfikatorów, zrezygnowano z jego użycia. Powodem rezygnacji było ryzyko wystąpienia błędów klasyfikacji wynikających z przeuczenia sieci w przypadku ewentualnej zmiany wielkości zbioru treningowego na etapie rozwoju oprogramowania. Dodatkowymi argumentami, które zadecydowały o wyborze klasyfikatora SVM były: niższa skuteczność minimalna dla sztucznej sieci neuronowej oraz brak powtarzalności wyników klasyfikacji za jej pomocą tych samych kształtów. W tym kontekście porównywalne średnie skuteczności rozpoznawania gestów dla klasyfikatorów obu typów przy akceptowalnych czasach treningu i walidacji maszyny wektorów nośnych przemawiały za wyborem tego klasyfikatora. Ze względów wydajnościowych, do implementacji maszyny wektorów nośnych w systemie, docelowo wybrano bibliotekę LIBSVM w wersji dla języka C++. Pozwoliło to dodatkowo obniżyć czas treningu i walidacji gestów w porównaniu z wersją dla języka Java, wykorzystywaną w systemie WEKA.

Powodem, dla którego jako metodę parametryzacji kształtów dłoni wybrano, wspomnianą w rozdziale 5.2.1 metodę PGH, było zapewnianie przez nią niezmiennego rozkładu cech niezależnie od kąta obrotu dłoni. Jest to zgodne z założeniem projektowym stanowiącym, że kąt obrotu nie powinien być czynnikiem separującym klasy gestów i nie powinien wpływać na skuteczność ich rozpoznawania. Dodatkowo metoda ta jest niewrażliwa na wielkość parametryzowanego kształtu. Dzięki temu zmiana lokalizacji użytkownika względem punktu optymalnego odsłuchu nie zaniża skuteczności rozpoznawania gestów.

Rozpoznawanie gestów dynamicznych zrealizowano w oparciu o system wnioskowania oparty na logice rozmytej. Logika rozmyta, jako dziedzina modelowana za pomocą metod o niskim koszcie obliczeniowym, pozwala na rozpoznawanie gestów dynamicznych bez zauważalnych opóźnień. Metody logiki rozmytej, jako metody nie-

deterministyczne, w naturalny sposób odzwierciedlają różne sposoby wykonania tego samego gestu przez człowieka. Możliwe jest zachowanie wysokiej skuteczności rozpoznawania w przypadku wykonania w ramach określonego gestu ruchu, którego trajektoria znacząco odbiega od trajektorii wzorcowej. Taka właściwość wpływa pozytywnie na wygodę obsługi systemu, umożliwiając użytkownikowi wykonywanie gestów w sposób nieobciążający stawów. Dodatkowo, zastosowanie logiki rozmytej pozwala na wykorzystanie informacji o prędkości ruchu jako jednego z czynników decydujących o przynależności gestu dynamicznego do określonej klasy. Te zagadnienia zostaną przedstawione szerzej w rozdziale 5.8.3, w którym scharakteryzowano sposób wykorzystania logiki rozmytej do rozpoznawania gestów oraz w rozdziale 6.3, w którym zbadano skuteczność detekcji gestów wspieranej logiką rozmytą i bez jej zastosowania.

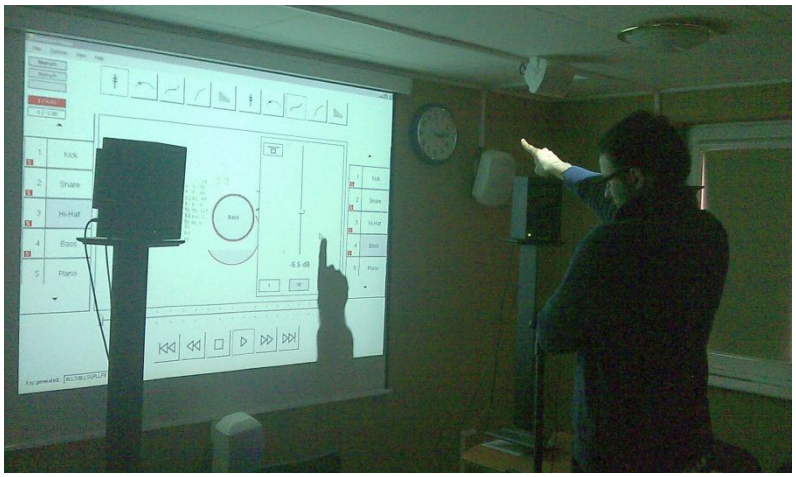
Sposób wykorzystania powyżej wyszczególnionych metod opisano w ramach kolejnego rozdziału rozprawy. W rozdziale tym przedstawiono wszelkie szczegóły dotyczące opracowania systemu miksowania dźwięku za pomocą gestów.

5.3 Komponenty i architektura systemu

Komponentami systemu są: komputer klasy PC, kamera internetowa, projektor multimedialny i ekran. Zastosowanie projektora multimedialnego i ekranu zamiast monitora wynikało z dwóch przyczyn. Pierwszą z nich było wymaganie dotyczące stworzenia warunków zapewniających większe „zanurzenie" inżyniera w procesie miksowania dźwięku w porównaniu ze znanymi sposobami wykorzystującymi sprzęt komputerowy i oprogramowanie. Zastosowanie projektora i ekranu, na którym wyświetlany jest obraz dużych rozmiarów oferuje większą swobodę ruchu niż zastosowanie monitora. Drugą przyczyną było przyjęte założenie dotyczące stworzenia możliwości zapewnienia warunków odsłuchowych, w których nie występuje zjawisko filtracji grzebieniowej. Autorski sposób rozpoznawania gestów, w konfiguracji wykorzystującej projektor multimedialny i kamerę skierowaną na ekran, jest przedmiotem zgłoszenia patentowego „System i sposób sterowania komputerem" o numerze P.390165. Sposób miksowania dźwięku za pomocą gestów rąk został dodatkowo zgłoszony jako rozwiązanie innowacyjne.

Usytuowanie użytkownika w stosunku do komponentów systemu przedstawiono na rys. 5.2. Użytkownik znajduje się w punkcie optymalnego odsłuchu, pomiędzy pro-

jektorem a ekranem. Projektor podwieszony jest pod sufitem na takiej wysokości i w takiej odległości od ekranu, aby wyciągnięte do góry ręce użytkownika rzucały na ekran cień. Kamera znajduje się przed użytkownikiem, a jej obiektyw skierowany jest na ekran i ustawiony w taki sposób, aby wyświetlany przez projektor obraz był możliwie największych rozmiarów w pozyskiwanym strumieniu wizyjnym i był w nim zawarty w całości.



Rys. 5.2 Komponenty systemu i lokalizacja użytkownika

Rozwiązanie bazuje na odejmowaniu obrazów pozyskanych z kamery od obrazów wyświetlanych za pomocą projektora multimedialnego. Proces odejmowania obrazów, z uwzględnieniem operacji mających na celu ich wzajemne dopasowanie, został szczegółowo opisany w dalszej części niniejszego rozdziału. Dlatego w tym miejscu ograniczono się jedynie do przedstawienia ogólnych zarysów funkcjonowania systemu. Wynik odejmowania jest przetwarzany i następnie poddawany analizie pod kątem rozpoznawania gestów. Odejmowanie wykonywane jest w przestrzeni barw RGB. System operuje na obrazach o rozmiarach 320 x 240 pikseli. Sposób działania systemu opisano szczegółowo w dalszych rozdziałach (rozdziały 5.7 i 5.8).

System rozpoznaje zarówno dynamiczne gesty rąk, tj. gesty bazujące na trajektorii ruchu, jak i gesty statyczne, polegające na formowaniu dłoni w określony kształt. Oba typy gestów są ze sobą powiązane, tj. wykonanie tego samego ruchu, ale z dłonią uformowaną w inny kształt, interpretowane jest jako dwa różne gesty. Dodatkowo określone gesty, wykonane w odpowiedniej kolejności, również reprezentują klasę gestu.

Architektura systemu została przedstawiona w postaci diagramu komponentów (rys. G.4) umieszczonego w sekcji G dodatków do rozprawy. W części programowej systemu wyróżnić można trzy aplikacje. Aplikacja odpowiedzialna za rozpoznawanie gestów wywołuje przypisane im akcje systemowe, polegające na emulacji wciśnięcia przycisku na klawiaturze lub przycisku myszy. Dodatkowo odzwierciedla ruch ręki w postaci ruchu kursora myszy na ekranie. Aplikacją, odbierającą te zdarzenia, jest dostosowany do obsługi za pomocą gestów interfejs graficzny, opisany w rozdziale 5.4. Przechwycenie zdarzenia powoduje wygenerowanie odpowiadającego mu kodu MIDI, wysyłanego do trzeciej z aplikacji, tj. systemu DAW. Obsługa wirtualnych portów MIDI realizowana jest w oparciu o darmowe oprogramowanie MIDI Yoke. Funkcje natywne, takie jak zmiana poziomu sygnału, rozpoczęcie odtwarzania czy ustawienie ścieżki w tryb samodzielnego odtwarzania (*solo*) obsługiwane są poprzez protokół mapujący MIDI HUI [137]. Protokół ten, ustalony przez firmy Mackie i DigiDesign w roku 1997, jest obecnie powszechnie wspierany przez systemy DAW i wykorzystywany do komunikacji z zewnętrznymi kontrolerami. Parametry wtyczek programowych spoza zbioru obsługiwanego przez protokół MIDI HUI powiązane są z poszczególnymi gestami poprzez komunikaty MIDI przypisane za pomocą funkcji *MIDI Learn*, będącej na wyposażeniu większości profesjonalnych systemów DAW. Parametrami tymi są: wzmocnienie korektora, próg i stopień kompresji oraz miks i czas pogłosu. Kontrolerom powyższych parametrów przypisano komunikaty kanałowe – głosowe, zwyczajowo odpowiedzialne za przesyłanie informacji o sile docisku klawisza (ang. *Aftertouch* / *Polyphonic key pressure*). Wartość siły docisku odpowiada w tym przypadku aktualnej wartości ustawianego parametru. Za pomocą komunikatów tego samego typu sterowane są również funkcje omijania (ang. *bypass*) korektora, kompresora dynamiki i pogłosu w torze sygnałowym. Pozostałe parametry sygnału, tj. poziom i panorama jako funkcje natywne, obsługiwane są poprzez protokół mapujący MIDI HUI. Zgodnie z tym protokołem ustawianie poziomu wybranego sygnału realizowane jest za pomocą komunikatu kanałowego – głosowego, odpowiedzialnego według standardowego protokołu MIDI za płynną zmianę wysokości dźwięku (ang. *Pitch wheel change*). Wartość poziomu reprezentowana jest domyślnie za pomocą jednego bajta, co pozwala na osiągnięcie 128 poziomów kwantyzacji. Korzystając z opcji w menu opracowanego interfejsu, użytkownik może jednak ustawić zapisywanie wartości na dwóch bajtach, zwiększając tym

samym dokładność do 16384 poziomów kwantyzacji. Panorama obsługiwana jest za pomocą komunikatu modyfikacji kontrolera (ang. *Control change*). Zgodnie ze specyfikacją protokołu MIDI HUI [137], w komunikacie, zamiast bezwzględnych wartości parametrów, jak ma to miejsce dla pozostałych parametrów, przesyłane są liczby 1 lub 65 odpowiednio dla przesunięcia źródła sygnału w bazie stereofonicznej o jedną jednostkę w prawo lub lewo.

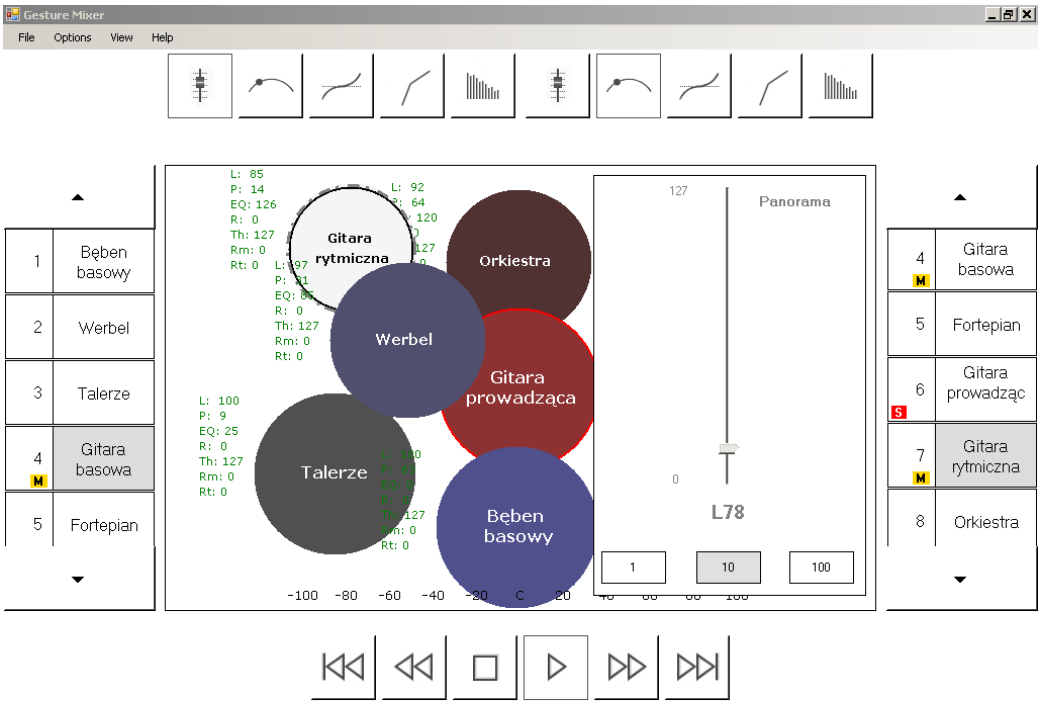
Obsługa edycji wartości parametrów za pomocą tylko jednej ręki realizowana jest poprzez emulowanie wciśnięcia lewego przycisku myszy, kiedy kursor znajduje się nad oknem edycji parametrów. Jak już wcześniej wspomniano, ruch ręki powoduje ruch kursora myszy na ekranie, co z kolei, przy emulowaniu wciśnięcia przycisku myszy, przekłada się na zmianę położenia suwaka w oknie edycji parametru. Możliwość jednoczesnej edycji dwóch parametrów zapewniona została w oparciu o komunikację opartą na gniazdach (ang. *sockets*) według modelu TCP (ang. *Transmission Control Protocol*). W momencie wykrycia wyprostowanych palców wskazujących obu dłoni i schowanych pozostałych, uruchamiane jest wysyłanie pozycji rąk z aplikacji rozpoznawania gestów do aplikacji interfejsu graficznego. Sposób wyznaczania pozycji rąk w obrazie opisano w punkcie 5.8.2. Wartości pozycji rąk lub kursora myszy tłumaczone są następnie na kody MIDI, zgodnie z opisem zawartym w powyższym paragrafie.

Dzięki podziałowi funkcjonalności rozpoznawania gestów i sterowania procesami miksowania dźwięku na dwie aplikacje, po skonfigurowaniu powiązań gest – akcja, możliwe jest obsługiwanie za pomocą gestów również innych aplikacji.

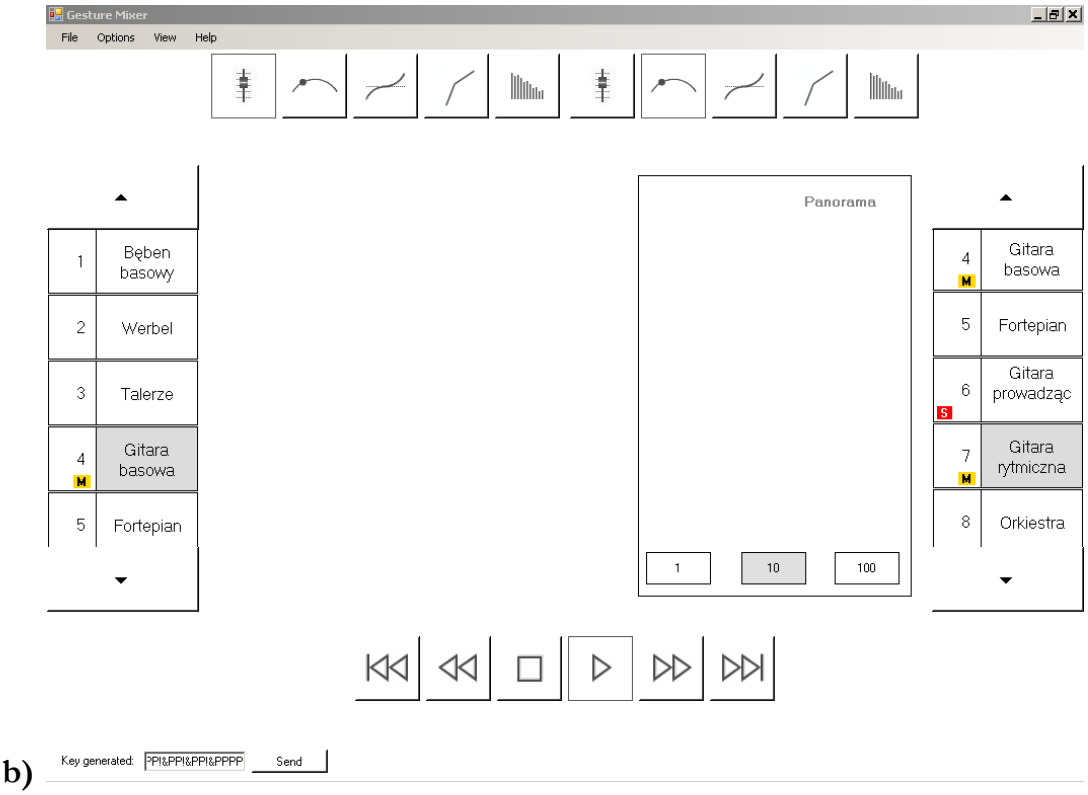
5.4 Interfejs graficzny

Zgodnie z założeniami przedstawionymi w rozdziale 5.1 opracowany interfejs stanowi nakładkę graficzną na interfejs wybranej aplikacji DAW, dostosowaną do obsługi za pomocą gestów. Aby możliwe było zbadanie wpływu reprezentowania wartości parametrów za pomocą informacji wizualnej na sposób miksowania, w systemie zaimplementowano dwa tryby interfejsu – pełny i ograniczony. W trybie pełnym interfejs zaprojektowano na podobieństwo systemu wizualizowania zgrań przedstawionego w filmie „*The art of mixing: a visual guide to recording engineering production*" [42]. Sygnały foniczne reprezentowane są w postaci kół (rys. 5.3a) osadzanych w różnych częściach ekranu. Pozycja koła na osi poziomej odpowiada lokalizacji źródła w bazie stereofo-

nicznej. Źródło, dla którego reprezentujący je kształt umieszczony jest na środku ekranu, zlokalizowane jest w centrum panoramy dźwiękowej. Wielkość kształtu odpowiada poziomowi dźwięku. Lokalizacja kształtu na osi pionowej reprezentuje wzmocnienie górno-zakresowego korektora półkowego z częstotliwością graniczną 7 kHz. Pozostałe parametry, tj. próg i stopień kompresji oraz czas pogłosu i stosunek dźwięku pogłosowego do bezpośredniego, zwany dalej miksem pogłosu, przedstawione są w formie liczbowej obok koła (rys. 5.3a). Dokładne wartości poziomu, panoramy i wzmocnienia korektora są także przedstawione w formie liczbowej.



a) Key generated: PI&PPI&PPI&PPI&PP Send



Rys. 5.3 Widok interfejsu graficznego w pełnym trybie graficznym (a) i trybie ograniczonym (b)

W częściach bocznych głównego okna interfejsu rozmieszczone są panele zawierające nazwy dostępnych źródeł sygnałów. Dla jednoczesnej modyfikacji dwóch parametrów sygnał wybrany z lewej strony interfejsu jest domyślnie skojarzony z lewą ręką, natomiast sygnał wybrany z prawej – z prawą. Możliwe jest wybranie tego samego sygnału dla lewej i prawej ręki, ale przypisanie innego parametru. Przykładowo, użytkownik może wykorzystać tę cechę do edycji ustawień kompresji dynamiki, będącej funkcją dwóch parametrów: stopnia i progu kompresji. W górnej części okna interfejsu znajdują się dwa panele (rys. 5.3) z identycznymi pięcioma piktogramami, odpowiedzialnymi za wybór do edycji kolejno następujących parametrów: poziomu, panoramy, wzmocnienia korektora, ustawień kompresora dynamiki, ustawień pogłosu. Panel w lewej połowie ekranu zawiera piktogramy parametrów dla źródła wybranego w lewym panelu źródeł sygnałów. Analogicznie, panel w prawej połowie zawiera piktogramy parametrów dla źródła wybranego w prawym panelu. Wyboru parametrów dokonuje się za pomocą gestów dynamicznych zgodnie z opisem przedstawionym w kolejnym podrozdziale (rozdz. 5.5) lub poprzez umieszczenie odpowiednio uformowanej dłoni nad pik-

togramem reprezentującym dany parametr. Wykonanie jednej z wyżej wymienionych akcji powoduje wyświetlenie okna edycji (rys. 5.3). W trybie pełnego interfejsu graficznego w oknie tym pojawia się suwak z wartością parametru u dołu (rys. 5.3a). Okno wyświetlane jest w lewej połowie ekranu, jeśli wybór parametru nastąpił w lewym górnym panelu i w prawej połowie, przy wyborze dokonanym w prawym panelu. Wybranie ustawień kompresora dynamiki lub pogłosu powoduje wyświetlenie dodatkowego panelu z dwoma przyciskami pod wciśniętym przyciskiem. Panel ten zawiera odpowiednio: przyciski odpowiedzialne za wybranie do edycji progu lub stopnia kompresji albo miksu lub czasu pogłosu. W trybie pełnego interfejsu edycja parametrów przypisanych do trzech pierwszych przycisków, tj. poziomu, panoramy i wzmocnienia korektora, może być realizowana bezpośrednio poprzez interakcję z kształtem reprezentującym źródło sygnału. Sposób interakcji został opisany w kolejnym podrozdziale (rozdz. 5.5). Wybranie kształtu przez osobę praworęczną powoduje zaznaczenie odpowiadającego źródła sygnału w panelu z prawej strony ekranu jako aktywnego (rys. 5.3a). Ustawienie w menu aplikacji opcji obsługi przez osobę leworęczną sprawia, że wybranie kształtu powoduje zaznaczenie źródła w panelu z lewej strony ekranu. Podobnie, dokonanie wyboru w jednym z paneli sygnałów, w zależności od wyboru trybu obsługi dla lewo albo praworęcznych, powoduje zaznaczenie odpowiadającego kształtu przez otoczenie go czarnym okręgiem (rys. 5.3a). W trybie ograniczonego interfejsu źródła foniczne nie posiadają reprezentacji graficznej, a edycji parametrów nie towarzyszy wyświetlanie suwaka z wartością (5.3b). W postaci graficznej dostępna jest jedynie informacja o tym, jaki parametr i dla którego źródła jest w danej chwili wybrany do edycji. W ten sposób, użytkownik systemu jest zmuszony podejmować wszelkie decyzje związane z miksowaniem jedynie na podstawie oceny zmiany dźwięku spowodowanej określonym ruchem ręki. Górne oraz boczne panele interfejsu wraz z informacjami o stanie systemu (odtwarzanie, zatrzymanie, przewijanie, itp.) oraz stanie źródeł (wyciszone / w trybie solo) są dostępne w obu trybach graficznych.

Po uruchomieniu systemu i zaimportowaniu sygnałów fonicznych obiekty reprezentujące źródła umieszczane są na środku ekranu i wszystkie przyjmują taki sam rozmiar, podobnie jak dla każdego kanału wirtualnego miksera w typowym oprogramowaniu DAW tłumiki domyślnie ustawione są w pozycji „0”, a pokrętła odpowiedzialne za regulację panoramy wskazują jej środek. Wyboru źródła, w przypadku gdy wszystkie

znajdują się jedno nad drugim, dokonuje się za pomocą panelu zawierającego nazwy źródeł z lewej lub prawej strony interfejsu (rys. 5.3). Z kolei, gdy źródła rozmieszczone są w różnych miejscach na ekranie i nie przesłaniają się wzajemnie, wyboru sygnałów do edycji można dokonywać za pomocą gestów. Sposób wybierania źródeł i edycji parametrów opisano w kolejnym podrozdziale (rozdz. 5.5).

5.5 Słownik gestów

W celu obsługi operacji miksowania dźwięku opracowano zunifikowany słownik gestów rąk. Słownik ten powstał w oparciu o badanie ankietowe oczekiwań potencjalnych użytkowników systemu względem sposobu obsługi za pomocą gestów oraz intuicyjności gestów i przypisanych im funkcji, zaproponowanych przez autora rozprawy. W badaniu udział wzięły 34 osoby, wśród których wyodrębniono trzy grupy o różnym stopniu znajomości aparatury studyjnej i procesów postprodukcji dźwięku. Formularz ankiety zamieszczono w sekcji A dodatków do rozprawy. Zadaniem ankietowanych było najpierw przypisanie do funkcji statycznych i dynamicznych gestów rąk wybranych z predefiniowanego zbioru. Następnie ankietowani byli proszeni o zaproponowanie własnych gestów, mogących wykraczać poza ramy określone zbiorem. Kolejno, ankietowani oceniali w pięciostopniowej skali intuicyjność powiązań gest – akcja, zaproponowanych przez autora rozprawy. Ponieważ wyniki wstępnych badań intuicyjności nie są bezpośrednio związane z kluczowymi wątkami niniejszej rozprawy, a jedynie pomogły w opracowaniu optymalnego słownika gestów, ze względu na obszerność rozprawy nie zostały one w niej przedstawione.

Ruchy ręką z dłonią skierowaną płaską stroną do ekranu nie mają przypisanej akcji. Dzięki temu możliwe jest intuicyjne wybieranie funkcji lub parametrów miksowania poprzez skierowanie dłoni nad ikonę na pasku menu. Do wykonania określonego gestu dłoń formuje się w odpowiedni kształt. Zestaw domyślnych gestów statycznych może być modyfikowany przez użytkownika na etapie trenowania klasyfikatorów. Do wykonania gestu dynamicznego realizującego wywołanie określonej operacji dłoń formuje się w kształt nr 3 (tabela 5.6). Powiązania gestów dynamicznych wykonywanych dla takiego kształtu dłoni z obsługiwanymi funkcjami przedstawiono w tabeli 5.7. Gesty dynamiczne przedstawiono w tabelach za pomocą strzałek określających trajektorie ruchu. Strzałki pojedyncze symbolizują gesty wykonywane jedną ręką, natomiast po-

dwójne – gesty obu rąk. Cyfra 2 w tabeli 5.7 oznacza, że gest wykonywany jest obiema rękoma. W trybie pełnego interfejsu graficznego, złożenie palców dłoni, umieszczonej nad obiektem graficznym reprezentującym źródło sygnału w kształt nr 2 (tabela 5.6) powoduje wybranie źródła do edycji. Za pomocą ruchu ręki z dłonią uformowaną w taki kształt możliwe jest przesuwanie obiektu w płaszczyźnie dwuwymiarowej, czyli zgodnie z opisem w poprzednim rozdziale, zmienianie lokalizacji źródła w panoramie lub modyfikowanie wzmocnienia korektora półkowego. Zakończenie przesuwania obiektu wykonywane jest przez uformowanie dłoni w kształt 1 (tabela 5.6). Źródło pozostaje jednak wybrane do edycji, tj. wykonanie jednego z gestów o numerach: 14 – 17 (tabela 5.7) lub umieszczenie dłoni uformowanej w kształt nr 1 (tabela 5.6) nad piktogramem parametru fonicznego powoduje wybranie parametru dla tego źródła. Dodatkowo, za pomocą wykonywanych obiema rękoma gestów nr 3 i 4 (tabela 5.7) możliwe jest odpowiednio zwiększanie i zmniejszanie obiektu, co odpowiada zwiększeniu bądź zmniejszeniu poziomu natężenia dźwięku. Po wybraniu parametru użytkownik może umieścić dłoń z palcami ułożonymi w gest nr 2 (tabela 5.6) nad wyświetlonym oknem edycji, co spowoduje pozyskanie aktualnej wartości parametru i uaktywnienie trybu edycji. Ruch ręki do góry powoduje zwiększanie wartości parametru, natomiast w dół – zmniejszanie. Uformowanie dłoni w kształt nr 1 (tabela 5.6) powoduje zakończenie edycji. W ten sposób możliwe jest również modyfikowanie panoramy i wzmocnienia korektora, dzięki czemu w trybie ograniczonego interfejsu użytkownik ma do dyspozycji ten sam zbiór funkcji, co w trybie pełnym. Dodatkowo, funkcjonalność taka stwarza możliwość jednoczesnego modyfikowania wartości dowolnych dwóch parametrów. Gesty o numerach: 14 – 17 (tabela 5.7), stanowiące alternatywę dla przywoływania parametrów poprzez umieszczenie dłoni nad piktogramem, zostały zdefiniowane w taki sposób, aby kojarzyły się semantycznie z realizowaną akcją. Przykładowo, wybranie progu kompresji do edycji wykonywane jest poprzez narysowanie litery "T" w powietrzu (od słowa *Threshold*).

Tabela 5.6 Domyślny zestaw gestów

1	✋	7	↓
2	☝	8	⇔
3	✌	9	⇒ ⇐
4	→	10	⇑ ⇓
5	←	11	⇓ ⇑
6	↑		

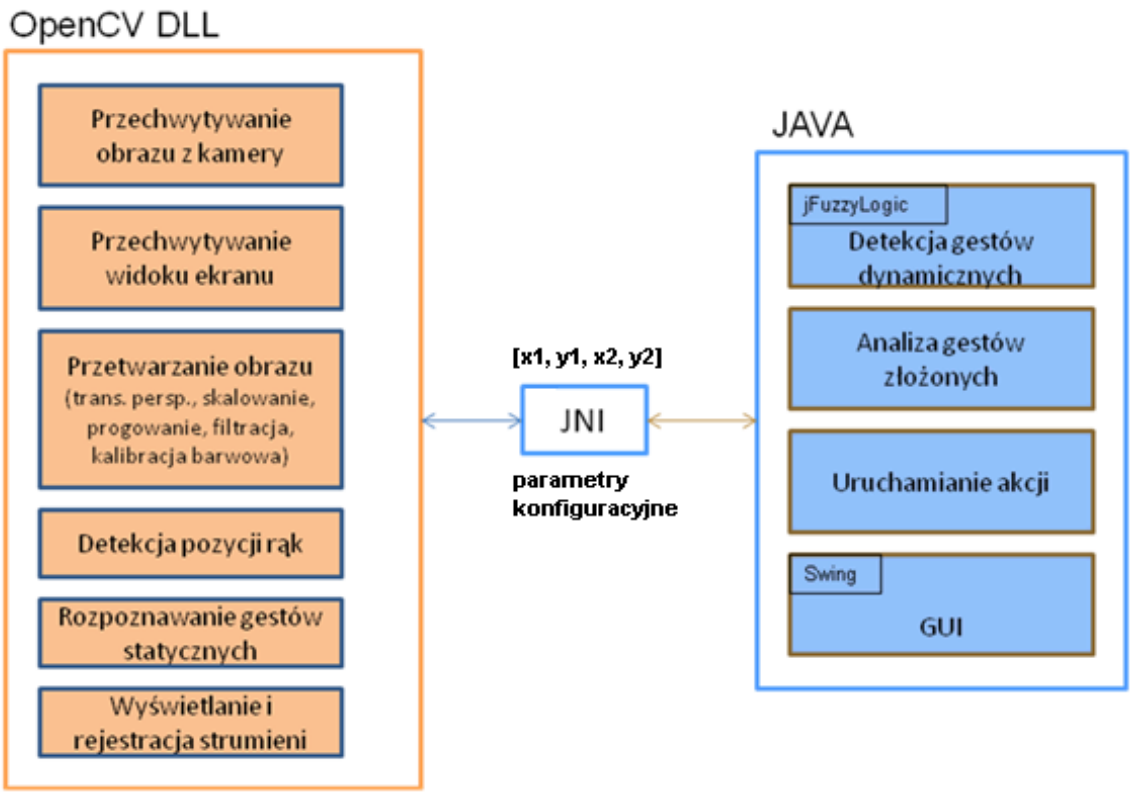
Tabela 5.7 Domyślne powiązania gestów i akcji

ID	Gest	Domyślna akcja
1	✋	Brak akcji
2	☝	Wybranie źródła sygnału
3	✋ ⇔	Zwiększenie poziomu sygnału
4	✋ ⇒ ⇐	Zmniejszenie poziomu sygnału
5	✌ → ←	Odtwarzanie
6	✌ ← →	Zatrzymanie
7	✌ → ←	Przewijanie do przodu (jeśli w trakcie odtwarzania)
8	✌ ← →	Przewijanie do tyłu (jeśli nie w trakcie odtwarzania)
9	✌↑↓	Solo / wyłączenie solo
10	✌↓↑	Wyciszenie / wyłączenie wyciszenia
11	$(✌↑↓)_2$	Wyłączenie trybu solo wszystkich ścieżek
12	$(✌↓↑)_2$	Wyłączenie trybu wyciszenia wszystkich ścieżek
13	☝ ↑ / ↓	Zwiększenie / zmniejszenie wartości wybranego parametru
14	✌ ⇓ ⇒	Wybranie czasu pogłosu do edycji
15	✌ → ⇓ ⇐ ↓	Wybranie stopnia kompresji dynamiki do edycji
16	✌ ⇒ ⇑	Wybranie progu kompresji dynamiki do edycji
17	✌ ⇓ ⇐ ⇒	Wybranie wzmocnienia korektora do edycji

5.6 Implementacja systemu

Do konstrukcji systemu wykorzystano część komponentów interfejsu rozpoznawania gestów rąk opracowanego przez autora rozprawy w ramach projektu *Typoszereg*, prowadzonego w Katedrze Systemów Multimedialnych. Aplikacja pierwotnego interfejsu została zaimplementowana w całości w oparciu o platformę Java SE (ang. *Java platform, Standard Edition*). Powodem wyboru tej technologii, poza spełnieniem wymagań dotyczących realizowalności postawionego problemu i wydajności, było dysponowanie komponentami, ocenionymi jako przydatne w rozwoju oprogramowania. Komponenty te zostały stworzone w ramach wcześniejszych prac nie związanych z prezentowanymi zagadnieniami. Dodatkowym atutem przemawiającym za zastosowaniem technologii Java SE była możliwość wykorzystania powszechnie stosowanego pakietu jFuzzyLogic, umożliwiającego interpretację gestów w oparciu o wnioskowanie rozmyte. Dla początkowo określonych zastosowań systemu, tj. zdalnego przeglądania obrazów i prezentacji multimedialnych, technologia ta pozwoliła osiągnąć satysfakcjonujący wynik w kontekście wydajności i skuteczności [85] [86]. Wydajność czasowa nie spełniała jednak określonych wymagań, biorąc pod uwagę zastosowanie systemu do obsługi tzw. „Wirtualnej Tablicy" [78] [79] [80] [84] czy wprowadzając możliwość interpretacji gestów zgrupowanych [83]. W takiej postaci system nie mógł być zastosowany do realizacji złożonych procesów towarzyszących miksowaniu dźwięku. Z tego względu wszelkie czasochłonne operacje, takie jak przetwarzanie klatek pozyskanych strumieni wizyjnych czy detekcja rąk w obrazie zaimplementowane zostały w środowisku Visual Studio C++ z wykorzystaniem biblioteki OpenCV (w wersji 2.2, w wariancie udostępnianym bez opłat). Podział funkcjonalności na platformy Java SE i C++ / OpenCV przedstawiono na rys. 5.4. Moduł interpretacji gestów dynamicznych pozostawiono zaimplementowany w języku JAVA z wykorzystaniem wspomnianego wcześniej pakietu jFuzzyLogic. Brak konieczności przenoszenia tej części systemu na platformę C++ / OpenCV stwierdzono przeprowadzając testy wydajnościowe, opisane w dalszej części rozprawy (rozdz. 6.2). W oparciu o platformę Java SE zaimplementowano również wywoływanie akcji systemowych przypisanych do gestów. Interfejs graficzny użytkownika został stworzony z wykorzystaniem pakietu Swing [107]. Funkcje napisane w języku C++ / OpenCV zostały udostępnione w postaci interfejsu JNI (ang. *Java Native*

Interface) i skompilowane do postaci biblioteki dynamicznej (*dll*) załączanej w kodzie Java. Aplikację interfejsu dedykowanego napisano w języku C++, wykorzystując technologię Windows Forms.



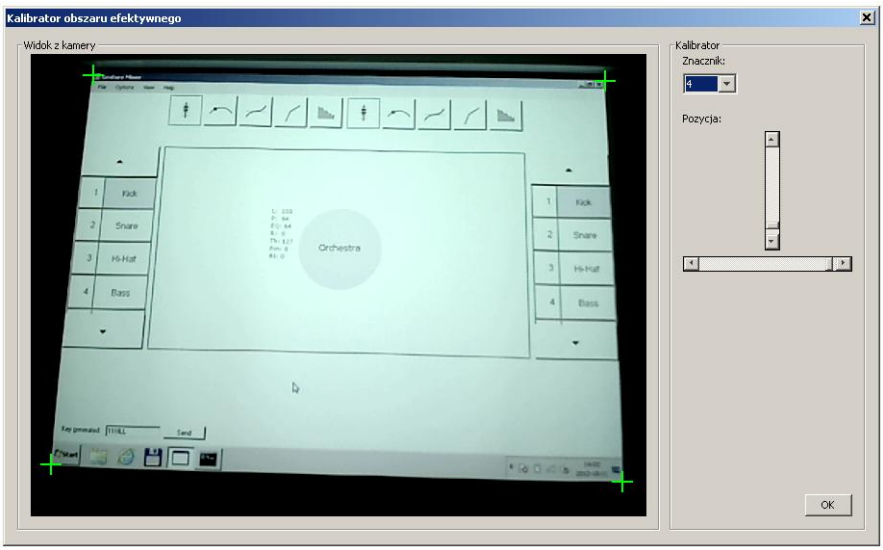
Rys. 5.4 Podział funkcji systemu na platformę Java SE i C++ / OpenCV

5.7 Uruchomienie i kalibracja systemu

Po zainstalowaniu pakietu oprogramowania system uruchamiany jest w tradycyjny sposób poprzez menu *Start* systemu operacyjnego lub poprzez kliknięcie ikony na pulpicie. Uruchomienie systemu powoduje automatyczną inicjalizację zarówno aplikacji odpowiedzialnej za rozpoznawanie gestów, jak i aplikacji dedykowanego interfejsu. W pierwszej kolejności uruchamiany jest dedykowany interfejs miksowania. Interfejs ten pełni rolę klienta w komunikacji za pomocą gniazd, oczekującego na komunikaty wysyłane przez aplikację serwera, tj. przez system rozpoznawania gestów. System DAW użytkownik uruchamia samodzielnie przed uruchomieniem wspomnianych aplikacji. Kolejność ta ma znaczenie, gdyż zainicjalizowanie interfejsu miksowania sprzężonego z systemem rozpoznawania gestów powoduje domyślnie wybranie pierwszej ścieżki w oknie programu DAW. Uruchomienie środowiska DAW po uruchomieniu interfejsu

miksowania mogłoby zatem prowadzić do niespójności w odzwierciedlaniu wyboru aktywnego źródła sygnału. Aby sterowanie parametrami środowiska DAW za pomocą gestów było możliwe, aplikacją o aktywnym oknie musi być aplikacja dedykowanego interfejsu.

Po zainicjalizowaniu aplikacji rozpoznawania gestów, na ekranie wyświetlane jest okno kalibratora obszaru pracy. Okno to zawiera klatkę strumienia wizyjnego pozyskanego z kamery z naniesionymi czterema krzyżykami (rys. 5.5). Zadaniem użytkownika jest wskazanie za ich pomocą rogów obrazu wyświetlanego przez projektor. Na podstawie wskazań obliczane są wartości zmiennych transformacji perspektywicznej wykonywanej w trakcie każdej iteracji rozpoznawania gestu. Określane są również wymiary obrazu wyświetlanego przez projektor, widocznego w klatce strumienia wizyjnego pobranego z kamery.



Rys. 5.5 Okno kalibratora obszaru efektywnego

Przed uruchomieniem procesu rozpoznawania gestów użytkownik może przeprowadzić proces kalibracji barwowej. Wykonanie kalibracji zwiększa skuteczność rozpoznawania gestów poprzez wyeliminowanie efektu winietowania wprowadzanego przez obiektyw kamery oraz redukcję zniekształceń spowodowanych nierównomiernym oświetleniem. W trakcie procesu kalibracji, dla określenia charakterystyki wpływu oświetlenia na dany kolor, wyświetlane są obrazy w całości wypełnione kolorem, kolejno czerwonym, zielonym, niebieskim, białym i czarnym. Dla każdego koloru tworzony jest profil korekcji barwowej. Proces tworzenia profili polega na odjęciu klatki strumie-

nia wizyjnego, złożonej z pikseli p_{ij}, od odpowiadającego obrazu wyświetlanego przez projektor, zgodnie z zależnościami (5.1–5.5). Odejmowanie realizowane jest w przestrzeni barw RGB. Piksele wyświetlanych obrazów przyjmują wartości z przestrzeni barw RGB równe odpowiednio p^{red} = [255, 0, 0], p^{green} = [0, 255, 0], p^{blue} = [0, 0, 255], p^{white} = [255, 255, 255] i p^{black} = [0, 0, 0]. Po odjęciu otrzymywane są profile korekcji składające się z pikseli p^{r}_{ij}, p^{g}_{ij}, p^{b}_{ij}, p^{wh}_{ij} i p^{bk}_{ij}, odpowiednio dla obrazów: czerwonego, zielonego, niebieskiego, białego i czarnego.

$$p_{ij}^{r} = \begin{cases} p^{red} - p_{ij} & \left| p^{red} - p_{ij} \geq 0 \right. \\ 0 & \left| p^{red} - p_{ij} < 0 \right. \end{cases} \tag{5.1}$$

$$p_{ij}^{g} = \begin{cases} p^{green} - p_{ij} & \left| p^{green} - p_{ij} \geq 0 \right. \\ 0 & \left| p^{green} - p_{ij} < 0 \right. \end{cases} \tag{5.2}$$

$$p_{ij}^{b} = \begin{cases} p^{blue} - p_{ij} & \left| p^{blue} - p_{ij} \geq 0 \right. \\ 0 & \left| p^{blue} - p_{ij} < 0 \right. \end{cases} \tag{5.3}$$

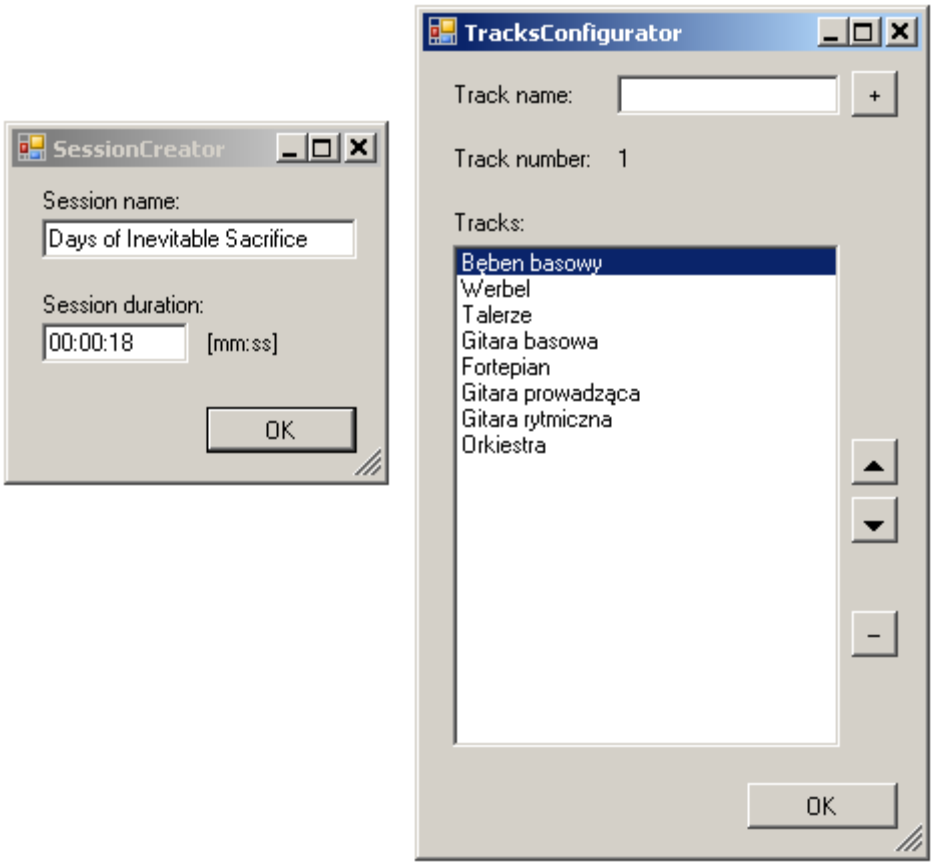
$$p_{ij}^{wh} = \begin{cases} p^{white} - p_{ij} & \left| p^{white} - p_{ij} \geq 0 \right. \\ 0 & \left| p^{white} - p_{ij} < 0 \right. \end{cases} \tag{5.4}$$

$$p_{ij}^{bk} = \left| p^{black} - p_{ij} \right| \tag{5.5}$$

Po wykonaniu kalibracji barwowej, kolejno w lewej i prawej części ekranu, wyświetlany jest prostokąt. Użytkownik umieszcza dłoń w takim miejscu w przestrzeni, aby rzucany przez nią cień powstawał w obrębie prostokąta, odpowiednio: w lewej części ekranu dla lewej ręki i prawej części dla ręki prawej. Proces ten ma na celu ustalenie przez użytkownika lokalizacji, przy której dłoń mieści się w całości we wskazanym obszarze. W późniejszych etapach trenowania klasyfikatorów i rozpoznawania gestów, analizowany jest fragment ręki określony wymiarami prostokąta. Przeprowadzenie tego procesu pozwala każdorazowo na usytuowanie użytkownika w punkcie optymalnego odsłuchu przy uwzględnieniu rozmieszczenia poszczególnych komponentów systemu zgodnie z zaleceniem z początku niniejszego rozdziału. Następnie przeprowadzany jest proces treningu zastosowanych w systemie klasyfikatorów gestów statycznych dla lewej

i prawej dłoni. Dla ustalonych warunków środowiskowych, tj. stałej lokalizacji systemu i względnie niezmiennego oświetlenia, proces ten dla danej osoby wykonywany jest tylko przy pierwszym uruchomieniu systemu, bezpośrednio po instalacji. Za rozpoznawanie gestów statycznych w systemie odpowiedzialne są maszyny wektorów nośnych (klasyfikatory SVM). Sposób wykorzystania klasyfikatorów oraz etap trenowania w kontekście podstaw teoretycznych metod przedstawiono odpowiednio w rozdziałach 5.2 i 5.8.4. W tym miejscu ograniczono się jedynie do opisu fazy trenowania z punktu widzenia użytkownika, tj. jako części etapu kalibracji systemu. Rozpatrywany w ten sposób proces trenowania klasyfikatorów jest identyczny ze sposobem ustalonym na potrzeby wykonania wstępnych badań klasyfikatorów opisanych w rozdziale 5.2. Użytkownik formuje dłoń w kształty należące do zbioru gestów statycznych (tabela 5.6) i porusza ręką kolejno, zgodnie z jedną z trzech trajektorii ruchu. Informacja o kształcie, w jaki w danej chwili należy uformować dłoń i trajektorii ruchu, znajduje się w górnej części ekranu. Dzięki możliwości wytrenowania systemu do rozpoznawania gestów zarówno jednej, jak i drugiej dłoni jednocześnie, użytkownik może z każdą dłonią skojarzyć wybrany parametr dla wybranego źródła sygnału i dokonywać modyfikacji dwóch parametrów w tym samym momencie.

Aby aplikacja interfejsu graficznego w prawidłowy sposób odzwierciedlała nazwy ścieżek w systemie DAW oraz ich liczbę i stan, użytkownik tworzy sesję w oknie konfiguratora (rys. 5.6) na wzór sesji w systemie DAW.



Rys. 5.6 Konfigurator sesji w aplikacji interfejsu graficznego

5.8 Zastosowane metody i algorytmy

5.8.1 Przetwarzanie wstępne obrazu

Sposób przetwarzania obrazu jest kluczowym czynnikiem decydującym o skuteczności detekcji gestów. Od zastosowanych metod zależy jakość reprezentacji poszukiwanych obiektów w obrazie, co z kolei przekłada się na skuteczność ich detekcji. Jak wspomniano na początku niniejszego rozdziału, zaproponowane rozwiązanie bazuje na odejmowaniu obrazów pozyskanych z kamery od obrazów wyświetlanych za pomocą projektora multimedialnego. Przed odjęciem, oba obrazy są przetwarzane w celu uzyskania podobnych charakterystyk. Przetwarzanie obrazu przebiega zgodnie z następującym schematem. Pozyskana z kamery klatka obrazu o rozmiarze 320 x 240 pikseli, zapisana w formacie BGR, poddawana jest transformacji perspektywicznej. Po jej wykonaniu z obrazu wycinany jest fragment zawierający jedynie obraz wyświetlany przez projektor. Fragment ten określony jest poprzez współrzędne jego wierzchołka oraz szerokość i wysokość. Wartości te wyznaczane są w oparciu o koordynaty punktów

naniesionych przez użytkownika na obraz z kamery podczas kalibracji obszaru pracy. Następnie, pozyskiwany jest obraz na wyjściu karty graficznej (wyświetlany przez projektor). Obraz ten odpowiada klatce obrazu pobranego z kamery. Zastosowana metoda zwraca obraz w formacie BGRA (ang. *blue*, *green*, *red*, *alpha*) o głębi koloru wynoszącej 8 bitów na kanał. Obraz ten poddawany jest przeskalowaniu w celu zapewnienia identycznych rozmiarów z przetworzonym obrazem z kamery i konwersji z formatu BGRA do formatu BGR o głębi koloru równej 8 bitów na kanał. Następnie, klatka obrazu pozyskana z kamery poddawana jest korekcji barwowej. Korekcja wykonywana jest w oparciu o przetworzony obraz na wyjściu karty graficznej i profile korekcji utworzone w procesie kalibracji barwowej, opisanej w rozdziale 5.7. Kolor każdego piksela p_{ij} obrazu z kamery, o wymiarach i x j, poddawany jest modyfikacji, zgodnie z zależnością 5.6, w której piksel p^c_{ij}, opisany zależnością 5.7, wchodzi w skład jednego z pięciu profili korekcji barwowej. Symbole r_{ij}, g_{ij}, b_{ij} w równaniu 5.7 oznaczają, odpowiednio, czerwoną, zieloną i niebieską składową piksela p_{ij}. Progi t^{rgb}, t^{wh} i t^{bk} zastosowano w celu rozróżniania, odpowiednio, pomiędzy komponentami RGB oraz białym i czarnym obrazem. Domyślne wartości progów zostały wyznaczone empirycznie i wyniosły, odpowiednio, 50, 180 i 80.

$$p'_{ij} = p_{ij} + p^c_{ij} \tag{5.6}$$

$$p^c_{ij} = \left\{ \begin{array}{l} p^r_{ij} \,|\, r_{ij} > g_{ij} + t^{rgb} \wedge r_{ij} > b_{ij} + t^{rgb} \\ p^g_{ij} \,|\, g_{ij} > r_{ij} + t^{rgb} \wedge g_{ij} > b_{ij} + t^{rgb} \\ p^b_{ij} \,|\, b_{ij} > r_{ij} + t^{rgb} \wedge b_{ij} > g_{ij} + t^{rgb} \\ p^{wh}_{ij} \,|\, r_{ij} > t^{wh} \wedge g_{ij} > t^{wh} \wedge b_{ij} > t^{wh} \\ -p^{bk}_{ij} \,|\, r_{ij} < t^{bk} \wedge g_{ij} < t^{bk} \wedge b_{ij} < t^{bk} \end{array} \right\} \tag{5.7}$$

Otrzymany obraz $\boldsymbol{p'}$ odejmowany jest od przetworzonego obrazu wyświetlanego przez projektor $\boldsymbol{p}^{screen}$, zgodnie z zależnością 5.8.

$$\mathbf{p}_{out} = \left| \mathbf{p}^{screen} - \mathbf{p}' \right| \tag{5.8}$$

Wynik odejmowania poddawany jest konwersji z przestrzeni barw RGB, przedstawionej w formacie BGR, do przestrzeni percepcyjnie ważonej skali szarości, zgodnie

z zależnością 5.9. W ten sposób otrzymywany jest obraz złożony z pikseli p_{ij}^{gray} o identycznych składowych RGB, oznaczonych odpowiednio jako r_{ij}^{gray}, g_{ij}^{gray}, b_{ij}^{gray}, zgodnie z zależnością 5.10.

$$r_{ij}^{gray} = g_{ij}^{gray} = b_{ij}^{gray} = \left[0.299 r_{ij}^{out} + 0.587 g_{ij}^{out} + 0.114 b_{ij}^{out}\right] \tag{5.9}$$

$$p_{ij}^{gray} = \left[r_{ij}^{gray}, g_{ij}^{gray}, b_{ij}^{gray}\right] \tag{5.10}$$

Otrzymany obraz jest binaryzowany z progiem domyślnie równym 100 zgodnie z zależnością 5.11. Wartość progu może być zmieniana przez użytkownika w panelu parametrów przetwarzania obrazu. Ostatnią operacją wykonywaną na obrazie jest filtracja medianowa. Przyjęto kwadratowy kształt maski, domyślnie o rozmiarze 7 x 7 pikseli.

$$p_{ij}^{bin} = \begin{cases} [0,0,0] & \left| r_{ij}^{gray} < 100 \right. \\ [255,255,255] & \left| r_{ij}^{gray} \geq 100 \right. \end{cases} \tag{5.11}$$

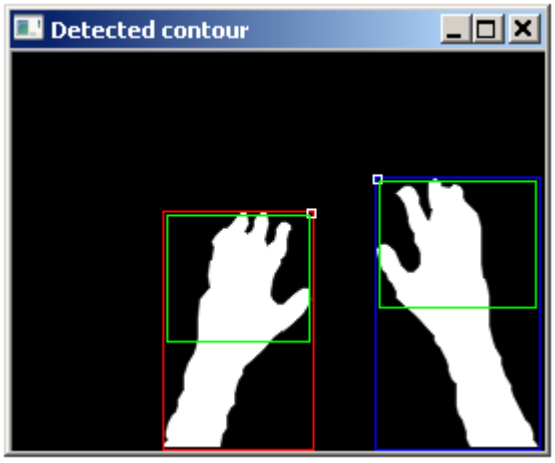
Przetworzony w ten sposób obraz, podawany jest na wejście algorytmu detekcji i śledzenia rąk opisanego w kolejnym podrozdziale.

5.8.2 Metoda detekcji i śledzenia rąk

Za detekcję rąk w obrazie odpowiedzialny jest algorytm z biblioteki OpenCV, bazujący na drzewach konturów – metodzie opracowanej przez Reeba [120] i dalej rozwiniętej przez Bajaja [5] i Carra [25]. Każdy obiekt w obrazie opisany jest za pomocą sekwencji punktów łączących odcinki tworzące jego kształt. Sekwencje te znajdowane są za pomocą metody Suzuki [134] i dla wyodrębnionego obiektu w obrazie zapisywane w formie drzewa. W drzewie tym korzeniem jest sekwencja reprezentująca zewnętrzny kontur a dziećmi, sekwencje opisujące wewnętrzne kontury kolejnych poziomów zagłębienia. Przykładowo dla dłoni przedstawiającej gest OK - liściem w drzewie - jest sekwencja opisująca wewnętrzny kontur kształtu utworzonego z kciuka i palca wskazującego. Spośród wszystkich obiektów znalezionych w obrazie jako ręce wybierane są te, których wysokość, szerokość i liczba pikseli je tworzących przekraczają określone progi. W ten sposób eliminowane są ewentualne szumy w obrazie, które nie zostały usunięte w procesie przetwarzania obrazu. Dla pierwotnego obrazu o wymiarach 320 x

240 pikseli, domyślne wartości progów liczby pikseli tworzących obraz ręki, szerokości i wysokości wyniosły odpowiednio: 100, 10 i 15. Wykryta ręka zaznaczana jest w obrazie za pomocą opisanego na niej prostokąta (rys. 5.7). Z takiego prostokątnego fragmentu obrazu wyodrębniony zostaje obraz samej dłoni (rys. 5.7). Współrzędne lewego, górnego wierzchołka obrazu dłoni i szerokość obrazu są identyczne ze współrzędnymi i szerokością prostokątnego obszaru całej ręki. Wysokość obrazu równa jest wysokości reprezentacji dłoni w obrazie, określonej w fazie kalibracji systemu, opisanej w rozdziale 5.7.

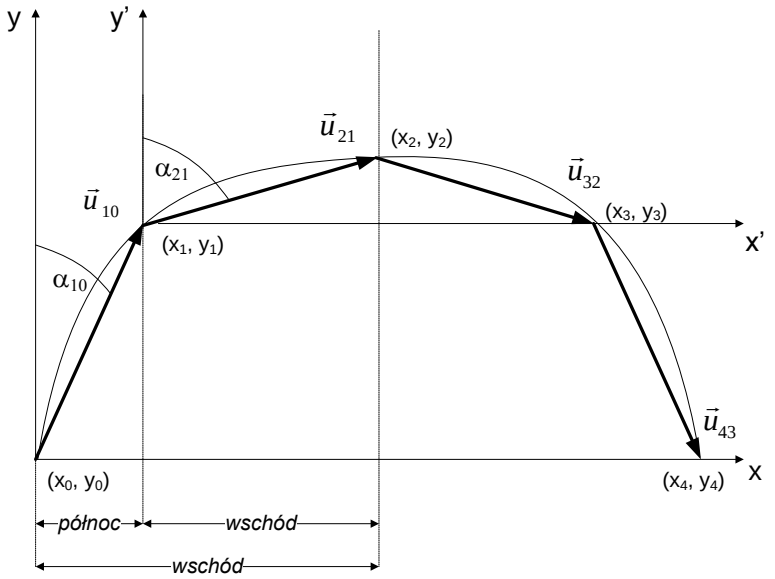
Sposób wyznaczania pozycji ręki zależny jest od fragmentu obrazu, jaki ona zajmuje. Jeśli ręka znajduje się w lewej połowie kadru, za pozycję przyjmuje się prawy górny róg prostokątnego obszaru, w który się ona wpisuje (rys. 5.7). Dla prawej połowy kadru pozycja ręki tożsama jest z lewym górnym rogiem tego obszaru. Przedstawiony sposób określania pozycji ręki w obrazie związany jest z przyjętą metodą rozpoznawania gestów dynamicznych, opisaną w kolejnym rozdziale i zastosowany został w celu wyeliminowania przypadku, w którym gest obu rąk w jego początkowej fazie interpretowany jest jako gest jednej ręki. Dla przykładu, rozważyć można sytuację, w której pozycje obu rąk określane są w taki sam sposób, tj. jako lewy górny róg obszaru dłoni. Wówczas w przypadku wykonywania gestu polegającego na przybliżeniu dłoni do siebie, jeśli ręce pierwotnie nie znajdowały się w kadrze, ruch w pierwszej kolejności zostanie sklasyfikowany jako przemieszczenie w prawo. Dopiero w momencie, gdy w kadrze pojawi się fragment lewej ręki o rozmiarze przekraczającym próg klasyfikacji obiektu, zostanie określona jej pozycja i następnie ruch zostanie poprawnie wykryty jako gest obu rąk. Przyjęty sposób wyznaczania pozycji rąk w obrazie wyklucza możliwość zaistnienia takiej sytuacji.



Rys. 5.7 Wykryte w obrazie ręce z zaznaczonymi punktami ich lokalizacji i obszarami dłoni (zielone ramki)

Przemieszczenie ręki zamodelowane zostało za pomocą wektorów ruchu. Pojedynczy wektor ruchu tworzony jest w oparciu o punkty stanowiące pozycje ręki określone dla klatek o numerach n i $n + 3$. Optymalny interwał pomiędzy kluczowymi ramkami, określony jako trzy klatki, zależny jest od rozdzielczości czasowej wyrażonej w liczbie przetwarzanych klatek obrazu w ciągu jednej sekundy. Został on dobrany w sposób empiryczny w oparciu o analizę wyników skuteczności detekcji dla sekwencji 20 strumieni wizyjnych zawierających gesty dynamiczne wykonywane przez 20 osób. Dobór interwału związany jest z przyjętą metodą rozpoznawania gestów dynamicznych, opartą na logice rozmytej, dlatego został wyjaśniony w kolejnym podrozdziale (rozdz. 5.8.3), związanym z tym zagadnieniem.

Każdy wektor $u_{ij} = [u^x_{ij}, u^y_{ij}]$ rozpatrywany jest pod względem prędkości i kierunku w kartezjańskim układzie współrzędnych (rys. 5.8).



Rys. 5.8 Wektory ruchu opisane na trajektorii ruchu ręki w prawo, z wyszczególnieniem początkowej fazy ruchu

Prędkość υ_{ij} wyznaczono zgodnie ze wzorem (5.12)

$$\upsilon_{ij} = \frac{\sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}}{t_i - t_j} \tag{5.12}$$

gdzie: $j = i - 1$

x_i – pozycja x ręki w obrazie w chwili t_i

x_j – pozycja x ręki w obrazie w chwili t_j

y_i – pozycja y ręki w obrazie w chwili t_i

y_j – pozycja y ręki w obrazie w chwili t_j

Kierunek wyrażono poprzez kąt φ_{ij} będący w relacji określonej równaniem 5.13, z kątem α_{ij} pomiędzy wektorem ruchu a wersorem osi y zgodnie z równaniem 5.14.

$$\varphi_{ij} = \begin{Bmatrix} \alpha_{ij} & ,u_{ij}^x \geq 0 \\ 360° - \alpha_{ij} & ,u_{ij}^x < 0 \end{Bmatrix} \tag{5.13}$$

$$\alpha_{ij} = \frac{180° \cdot \arccos \dfrac{u_{ij}^y}{\left|\vec{u}_{ij}\right|}}{\pi} \; [°] \tag{5.14}$$

gdzie: α_{ij} – kąt pomiędzy wektorem u_{ij} a wersorem osi y

$\vec{u}_{ij}$ – wektor ruchu utworzony w oparciu o punkty w chwili t_i i t_{i-1}

u_y^{ij} – współrzędna y wektora utworzonego w oparciu o punkty w chwili t_i i t_{i-1}

u_x^{ij} – współrzędna x wektora utworzonego w oparciu o punkty w chwili t_i i t_{i-1}

W celu zapewnienia wysokiej skuteczności śledzenia rąk w obrazie, trajektorie ruchu są wygładzane za pomocą filtrów Kalmana [69]. Wpływ zastosowania filtracji Kalmana na skuteczność śledzenia rąk w obrazie zbadano i opisano w jednej z publikacji autora [81].

Do implementacji filtrów wykorzystano bibliotekę OpenCV [17]. Stanem systemu w chwili t, zgodnie z równaniem 5.15, jest wektor zawierający współrzędne x_t, y_t ręki w obrazie oraz prędkość poziomą i pionową wyrażone wzorami: 5.16 i 5.17.

$$\mathbf{s}_t = \left[x_t, y_t, \upsilon_t^x, \upsilon_t^y\right] \tag{5.15}$$

$$\upsilon_t^x = \upsilon_t \sin\varphi \tag{5.16}$$

$$\upsilon_t^y = \upsilon_t \cos\varphi \tag{5.17}$$

Zaimplementowany filtr Kalmana w bibliotece OpenCV, w ogólnej postaci, na podstawie obserwacji z chwili $t-1$ dokonuje predykcji stanu $\hat{\mathbf{s}}_{t|t-1}$ w chwili t zgodnie z równaniem 5.18.

$$\hat{\mathbf{s}}_{t|t-1} = \mathbf{F}_t \hat{\mathbf{s}}_{t-1|t-1} + \mathbf{B}\mathbf{u}_t + \mathbf{w}_t \tag{5.18}$$

gdzie $\mathbf{F}_t$ oznacza macierz przejścia stanu, $\hat{\mathbf{s}}_{t-1|t-1}$ jest stanem w chwili $t-1$, $\mathbf{B}$ jest opcjonalną macierzą, która koreluje ze zmianą stanu wektor $\boldsymbol{u}_t$, zawierający parametry umożliwiające sprawowanie zewnętrznej kontroli nad modelowanym systemem, a $\boldsymbol{w}_t$ oznacza wektor zakłócenia procesu. Ponieważ nie ma możliwości dostarczenia do opracowanego systemu wiedzy o trajektorii ruchu, równanie 5.18 upraszcza się do postaci 5.19.

$$\hat{\mathbf{s}}_{t|t-1} = \mathbf{F}_t \hat{\mathbf{s}}_{t-1|t-1} + \mathbf{w}_t \tag{5.19}$$

Wektor pomiarowy $\mathbf{z}_t$ (obserwacja) w chwili t dla rzeczywistego stanu chwilowego x_t przyjmuje postać 5.20.

$$\mathbf{z}_t = \mathbf{H}_t \mathbf{s}_t + \mathbf{v}_t \tag{5.20}$$

We wzorze 5.20 $\mathbf{H}_t$ jest modelem obserwacji mapującym rzeczywistą przestrzeń stanów na przestrzeń obserwowaną i dla opracowanego systemu przyjmuje postać wektora jednoelementowego o wartości 1, a $\mathbf{v}_t$ to wektor szumu pomiarowego. Estymata stanu *a posteriori* przyjmuje postać 5.21.

$$\hat{\mathbf{s}}_{t|t} = \hat{\mathbf{s}}_{t|t-1} + \mathbf{K}_t \cdot \left(\mathbf{z}_t - \mathbf{H}_t \hat{\mathbf{s}}_{t|t-1}\right) \tag{5.21}$$

Macierz $\mathbf{K}_t$ w równaniu 5.21 oznacza wzmocnienie Kalmana, określone wzorem 5.22.

$$\mathbf{K}_t = \mathbf{P}_{t|t-1} \mathbf{H}_t^{\mathrm{T}} \mathbf{S}_t^{-1} \tag{5.22}$$

Macierz $\mathbf{P}_{t|t-1}$ w równaniu 5.22 jest estymatą kowariancji *a priori*, określona równaniem 5.24, a $\mathbf{S}_t$ jest macierzą kowariancji rezydualnej (innowacji) określoną równaniem 5.23.

$$\mathbf{S}_t = \mathbf{H}_t \mathbf{P}_{t|t-1} \mathbf{H}_t^{\mathrm{T}} + \mathbf{R}_t \tag{5.23}$$

$$\mathbf{P}_{t|t-1} = \mathbf{F}_t \mathbf{P}_{t-1|t-1} \mathbf{F}_t^{\mathrm{T}} + \mathbf{Q}_t \tag{5.24}$$

Wartości wektora szumu przetwarzania $\mathbf{w}_t$ pozwalają na uwzględnienie w modelowanym systemie faktu, że prędkość ruchu nie jest stała. Wartości na przekątnej macierzy kowariancji szumu przetwarzania $\mathbf{Q}_t$ wyniosły 10^{-5}. Wartości wektora szumu pomiarowego $\mathbf{v}_t$ określono przyjmując dopuszczalny błąd pomiaru pozycji ręki w obrazie rzędu jednego piksela. W ten sposób, dla zastosowanej rozdzielczości obrazu 320 x 240 pikseli, macierz kowariancji szumu pomiarowego $\mathbf{R}_t$ przyjęła na przekątnej wartości równe $1 / (320 \cdot 240)$.

Stan w chwili t jest powiązany ze stanem w chwili $t - 1$ funkcją prędkości, zatem macierz przejścia przyjmuje postać 5.25.

$$\mathbf{F}_t = \begin{bmatrix} 1 & 0 & dt & 0 \\ 0 & 1 & 0 & dt \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.25}$$

Wielkość *dt* w macierzy $\mathbf{F}_t$, określona równaniem 5.26, oznacza modyfikator czasowy prędkości zależny od rozdzielczości czasowej f_{FR} systemu mierzonej w liczbie przetwarzanych klatek w ciągu jednej sekundy i wcześniej opisanego interwału $n_{\tau_0}^{\tau_1}$ pomiędzy kluczowymi ramkami. Zmienna *c* oznacza współczynnik skalowania wynikający z przyjętej jednostki prędkości i wynosi 10.

$$dt = c \cdot \frac{n_{\tau_0}^{\tau_1}}{f_{FR}} \tag{5.26}$$

Stosując macierz przejścia dla stanu w chwili $t - 1$, otrzymuje się przewidywany stan w chwili *t*, określony macierzą 5.27.

$$\hat{\mathbf{s}}_{t|t-1} = \begin{bmatrix} x_{t|t-1} = x_{t-1|t-1} + c \cdot \dfrac{n_{\tau_0}^{\tau_1}}{f_{FR}} \cdot \upsilon_{t-1|t-1}^{x} \\ y_{t|t-1} = y_{t-1|t-1} + c \cdot \dfrac{n_{\tau_0}^{\tau_1}}{f_{FR}} \cdot \upsilon_{t-1|t-1}^{y} \\ \upsilon_{t|t-1}^{x} = \upsilon_{t-1|t-1}^{x} \\ \upsilon_{t|t-1}^{y} = \upsilon_{t-1|t-1}^{y} \end{bmatrix} \tag{5.27}$$

5.8.3 Metoda rozpoznawania gestów dynamicznych

Gesty dynamiczne rozpoznawane są w oparciu o metody logiki rozmytej. Trajektorie ruchu rąk zamodelowano za pomocą 30 reguł rozmytych opisujących relacje pomiędzy kolejnymi wektorami ruchu. Jako pojęcia lingwistyczne opisujące wektor ruchu wybrano prędkość (ang. *velocity*) i kierunek (ang. *direction*). Ponieważ trajektorie ruchu analizowane są w ramach segmentów dwu-wektorowych dla lewej i prawej ręki, łącznie wyróżnić można osiem zmiennych lingwistycznych przedstawionych w tabelach 5.8 i 5.9. Trajektorie ruchu analizowano w kartezjańskim układzie współrzędnych (rys. 5.7).

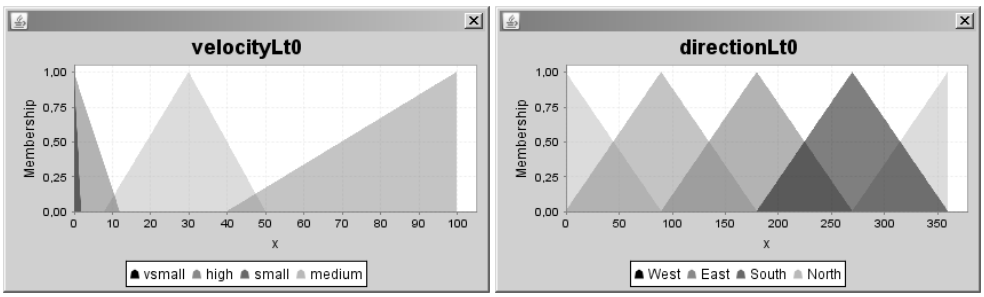
Tabela 5.8 Zmienne lingwistyczne dla prędkości ruchu ręki

symbol prędkości	interwał	opis
υ_{21}^{L}	$t_2 - t_1$	prędkość ruchu lewej ręki
υ_{10}^{L}	$t_1 - t_0$	
υ_{21}^{R}	$t_2 - t_1$	prędkość ruchu prawej ręki
υ_{10}^{R}	$t_1 - t_0$	

Tabela 5.9 Zmienne lingwistyczne dla kierunku ruchu ręki

symbol kierunku	interwał	opis
α_{21}^{L}	$t_2 - t_1$	kierunek ruchu lewej ręki
α_{10}^{L}	$t_1 - t_0$	
α_{21}^{R}	$t_2 - t_1$	kierunek ruchu prawej ręki
α_{10}^{R}	$t_1 - t_0$	

Zdefiniowano cztery zbiory rozmyte dla wartości prędkości i cztery zbiory rozmyte dla kątów określających kierunek (rys. 5.9). Zbiory rozmyte dla prędkości określono za pomocą pojęć lingwistycznych: *bardzo mała* (*vsmall*), *mała* (ang. *small*), *średnia* (ang. *medium*) i *duża* (ang. *high*). Zbiory rozmyte dla kierunków zostały określone poprzez pojęcia: *północ* (ang. *North*), *wschód* (ang. *East*), *południe* (ang. *South*), *zachód* (ang. *West*). Powyższe oznaczenia w kodzie FCL (ang. *Fuzzy Control Language*), przedstawionym w dalszej części rozprawy i na rys. 5.9, wygenerowanym przez moduł wnioskowania opisany tym językiem, przyjęto w języku angielskim mając na względzie dobre praktyki programowania. Zbiór rozmyty dla kierunku *północ* zdefiniowano za pomocą dwóch funkcji trójkątnych, wyznaczających przedziały [0°, 90°] i [270°, 360°].



Rys. 5.9 Zbiory rozmyte dla zmiennych lingwistycznych: *prędkość* i *kierunek*

Dla wszystkich zbiorów przyjęto trójkątne funkcje przynależności. Zbiory wynikowe każdej reguły, reprezentujące klasy gestów, przyjęły postać singletonów, zgodnie z modelem Takagi-Sugeno zerowego rzędu [135]. Wartością wynikową systemu wnioskowania była odpowiedź reguły, dla której wartość wynikowej funkcji przynależności była maksymalna. Dodatkowo wprowadzono próg o wartości 0,5, poniżej którego aktywność ruchowa nie była kojarzona z żadnym z gestów. W ten sposób rozwiązano problem mylnego przypisywania znaczenia ruchom przejściowym pomiędzy kolejnymi gestami.

W celu uwzględnienia naturalności motoryki ludzkich gestów reguły rozmyte zostały zdefiniowane w oparciu o wspomnianą sekwencję 20 strumieni wizyjnych zawierających zarejestrowane ruchy. Zauważono, że pełen ruch ręką przykładowo z lewej do prawej strony kadru często wykonywany jest po okręgu. Po przedstawieniu takiego ruchu w postaci trajektorii opisanej wektorami (rys. 5.8) wyróżnić można kierunki, dla których wartości funkcji przynależności należą w największym stopniu do zbiorów rozmytych oznaczonych kolejno jako: *północ*, *wschód*, *południe*. Ze względu na wyszczególnienie trzech faz ruchu przemieszczenie ręki w prawo zostało zamodelowane za pomocą trzech reguł rozmytych, zgodnie z kodem FCL przedstawionym w dalszej części tego rozdziału. Zgodnie z przyjętym modelem trajektoria takiego ruchu powinna zostać oznaczona przez system za pomocą wektorów, spośród których nie więcej niż jeden ma kierunek *północ* i nie więcej niż jeden – kierunek *południe*. Warunkiem takiego rozkładu wektorów jest optymalny wybór wspomnianego w poprzednim podrozdziale interwału pomiędzy ramkami, w oparciu o które tworzone są wektory ruchu. Ustalono, że interwał trzech klatek odpowiadający czasowi wynoszącemu średnio 182 ms, przy uzyskanych 22 klatkach na sekundę, pozwalał na prawidłowe rozpoznawanie gestów 18 z 20 osób biorących udział w testach. W przypadku skrócenia tego czasu, gdy użytkownik wykonywał powolny ruch, początkowa faza wykonywania gestu reprezentowana była przez dwa wektory o kierunku *północ*, co powodowało sklasyfikowanie gestu jako ruch do góry, zamiast w prawo. Z kolei, rzadsze próbkowanie ramek, przy szybkim ruchu powodowało niekiedy jego opisanie za pomocą tylko jednego wektora. W przypadku, gdy wektor ten miał kierunek *północ* lub *południe* gest był rozpoznawany nieprawidłowo. Dla optymalnego interwału pomiędzy kluczowymi ramkami problem ten nie występował. Zauważyć należy, że aby omawiany w ramach tego przykładu gest

był rozpoznany prawidłowo, trajektoria ruchu może przyjmować inne postaci, które są podzbiorem postaci przedstawionej na rys. 5.8.

Przyjęta metoda rozpoznawania gestów dynamicznych w sposób naturalny odzwierciedla proces śledzenia i nazywania ruchów przez człowieka. Zastosowana metoda pozwala na rozpoznawanie gestów i podejmowanie akcji w sposób ciągły. Innymi słowy, nie jest konieczne pozyskanie pełnej informacji o ruchu do prawidłowego wnioskowania o tym, jaki gest ten ruch reprezentuje. Podejście takie cechuje również zdecydowanie szybszy czas reakcji systemu na wykonany ruch niż w przypadku podejścia, w którym do wnioskowania o geście konieczne jest pozyskanie pełnej trajektorii ruchu.

Poniżej przedstawiono przykładowe reguły rozmyte dla analizowanego powyżej gestu jednej ręki i gestu obu rąk. Pierwsze trzy reguły, określające gest jednej ręki, dotyczą ruchu wykonywanego jedynie lewą ręką, dlatego z uwagi na przyjęty sposób implementacji prędkość prawej ręki powinna zawierać się w zbiorze rozmytym *bardzo mała*. Jako przykład zamodelowania gestu obu rąk poniżej przedstawiono również regułę reprezentującą aktywność ruchową polegającą na oddalaniu jedna od drugiej, wyciągniętych przed siebie rąk. Na podstawie analizy materiałów wizyjnych z zarejestrowanymi sekwencjami gestów użytkowników ustalono, że naturalnym sposobem wykonywania takiego ruchu jest wzajemne oddalanie jednej ręki od drugiej z zachowaniem stałego kierunku. Z tego względu w przedstawionej regule rozmytej wartość zmiennych lingwistycznych *kierunek* należy do zbioru rozmytego *zachód* dla obu wektorów ruchu prawej ręki i do zbioru rozmytego *wschód* dla obu wektorów ruchu lewej ręki. W trakcie wykonywania gestu obu rąk prędkość ruchu może być mniejsza niż w przypadku gestów tylko jednej ręki. Z tego względu, w przeciwieństwie do pierwszych trzech reguł, w czwartej przedział prędkości decydujący o sklasyfikowaniu ruchu jako określonego gestu rozszerzono o wartości zbioru *mała*.

$$\forall\begin{pmatrix}\varphi_{10}^{L}\in N\wedge\varphi_{21}^{L}\in E\wedge\upsilon_{10}^{L}\notin S\wedge\\ \upsilon_{21}^{L}\notin S\wedge\upsilon_{10}^{R}\in VS\wedge\upsilon_{21}^{R}\in VS\end{pmatrix}\Rightarrow \\ \Rightarrow g\in G_2 \tag{5.28}$$

$$\forall\begin{pmatrix}\varphi_{10}^{L}\in E\wedge\varphi_{21}^{L}\in E\wedge\upsilon_{10}^{L}\notin S\wedge\\ \upsilon_{21}^{L}\notin S\wedge\upsilon_{10}^{R}\in VS\wedge\upsilon_{21}^{R}\in VS\end{pmatrix}\Rightarrow \\ \Rightarrow g\in G_2 \tag{5.29}$$

$$\forall\begin{pmatrix}\varphi_{10}^{L}\in E\wedge\varphi_{21}^{L}\in S\wedge\upsilon_{10}^{L}\notin S\wedge\\ \upsilon_{21}^{L}\notin S\wedge\upsilon_{10}^{R}\in VS\wedge\upsilon_{21}^{R}\in VS\end{pmatrix}\Rightarrow \\ \Rightarrow g\in G_2 \tag{5.30}$$

$$\forall\begin{pmatrix}\varphi_{10}^{L}\in W\wedge\varphi_{21}^{L}\in W\wedge\\ \varphi_{10}^{R}\in E\wedge\varphi_{21}^{R}\in E\wedge\upsilon_{10}^{L}\notin VS\wedge\\ \upsilon_{21}^{L}\notin VS\wedge\upsilon_{10}^{R}\notin VS\wedge\upsilon_{21}^{R}\notin VS\end{pmatrix}\Rightarrow \\ \Rightarrow g\in G_6 \tag{5.31}$$

Poniżej przedstawiono reprezentację reguł 5.28 – 5.31 w postaci kodu języka FCL. W części B dodatków zamieszczono w postaci kodu FCL wszystkie reguły, zawarte wraz z pozostałymi elementami opisu systemu wnioskowani w pliku .fcl, z którego korzysta opracowany system.

```
// początkowa faza ruchu lewej ręki w prawo
RULE 8 : IF directionLt0 IS North AND directionLt1 IS East
      AND velocityLt0 IS NOT small AND velocityLt1 IS NOT small
      AND velocityRt0 IS vsmall AND velocityRt1 IS vsmall
      THEN gesture IS g2;

// środkowa faza ruchu lewej ręki w prawo
RULE 9 : IF directionLt0 IS East AND directionLt1 IS East
      AND velocityLt0 IS NOT small AND velocityLt1 IS NOT small
      AND velocityRt0 IS vsmall AND velocityRt1 IS vsmall
      THEN gesture IS g2;

// końcowa faza ruchu lewej ręki w prawo
RULE 10 : IF directionLt0 IS East AND directionLt1 IS South
      AND velocityLt0 IS NOT small AND velocityLt1 IS NOT small
      AND velocityRt0 IS vsmall AND velocityRt1 IS vsmall
      THEN gesture IS g2;

// oddalenie rąk, jedna od drugiej
RULE 26 : IF directionLt0 IS West AND directionLt1 IS West
      AND directionRt0 IS East AND directionRt1 IS East
      AND velocityLt0 IS NOT vsmall AND velocityLt1 IS NOT vsmall
      AND velocityRt0 IS NOT vsmall AND velocityRt1 IS NOT vsmall
      THEN gesture IS g6;
```

Podkreślić należy, że przedstawiona metoda rozpoznawania gestów dynamicznych dotyczy przypadków, w których podstawową informacją analizowaną przy podejmowaniu decyzji o wywołaniu akcji przypisanej gestowi jest zmiana kierunku ruchu. Akcja przypisana do gestu nr 13 w tabeli 5.7, w rozdziale 5.5, nie jest skutkiem odpowiedzi przedstawionego powyżej systemu rozpoznawania gestów. Ruch towarzyszący

temu gestowi nie ma bowiem cech trajektorii separującej klasy gestów. O jego znaczeniu jednoznacznie decyduje wykonywany jednocześnie gest statyczny nr 2 z tabeli 5.6. oraz obecność okna edycji parametru na ekranie. Do wnioskowania o geście wykorzystano zatem w tym przypadku również chwilowy stan elementów interfejsu graficznego aplikacji. Jak już wspomniano wcześniej w rozdziale 5.5, wyprostowanie palca wskazującego i schowanie pozostałych palców powoduje emulowanie wciśnięcia przycisku myszy lub ciągłe wysyłanie pozycji dłoni poprzez kanały komunikacji za pomocą gniazd. Ponieważ w momencie emulowania wciśnięcia przycisku myszy kursor znajduje się nad oknem edycji z aktywną kontrolką w postaci suwaka, edycja parametru odbywa się tak samo, jak w przypadku użycia myszy. Wysyłanie pozycji przez kanały komunikacji za pomocą gniazd symuluje natomiast funkcjonalność myszy w sytuacji jednoczesnej edycji dwóch parametrów. Powyższy sposób podejmowania akcji dla gestu nr 13 został wybrany w celu zapewnienia edycji parametrów bez zauważalnych opóźnień, a tym samym z dużą dokładnością. Należy zauważyć, że przy zastosowaniu takiej samej metody rozpoznawania gestów i podejmowania akcji, jak dla pozostałych ruchów, zmiana wartości parametru mogłaby nastąpić dopiero po analizie sześciu ramek obrazu, zawierających dwa wektory ruchu. Wówczas rozpoznanie przemieszczenia ręki w górę odpowiadałoby zwiększeniu wartości parametru o jedną jednostkę, natomiast rozpoznanie przemieszczenia w dół – zmniejszeniu wartości. Taki model interakcji mógłby być postrzegany przez użytkownika jako skokowy i powolny, dlatego został odrzucony w trakcie projektowania systemu.

5.8.4 Metoda rozpoznawania gestów statycznych

Jak napisano w rozdziale 5.2.3 za rozpoznawanie gestów statycznych w systemie odpowiedzialne są maszyny wektorów nośnych (klasyfikatory SVM) typu C-SVC (ang. *C-Support Vector Classification*) [26]. Klasyfikatory zaimplementowano wykorzystując bibliotekę LIBSVM [26]. Wyboru metody dokonano w oparciu o przegląd badań w dziedzinie automatycznej klasyfikacji obiektów i zdarzeń i wstępne eksperymenty przeprowadzone przez autora rozprawy (rozdz. 5.2).

Wybrana metoda klasyfikacji sprowadza się do rozwiązania problemu optymalizacyjnego I rzędu wyrażonego postacią 5.31 przy warunku określonym zależnością 5.32.

$$\min_{w,b,\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \tag{5.31}$$

$$\begin{aligned} & y_i\left(\mathbf{w}^T\phi(\mathbf{x}_i)+b\right) \geq 1-\xi_i \\ & \xi_i \geq 0,\ i=1,..,l \end{aligned} \tag{5.32}$$

Formę dualną postaci 5.31 określa wzór 5.33, z warunkiem wyrażonym równaniem 5.34.

$$\min_{\alpha} \quad \frac{1}{2}\boldsymbol{\alpha}^T Q\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha} \tag{5.33}$$

$$\begin{aligned} & \mathbf{y}^T\boldsymbol{\alpha} = 0 \\ & 0 \leq \alpha_i \leq C,\ i=1,..,l \end{aligned} \tag{5.34}$$

W wyrażeniach 5.31 – 5.34, $\boldsymbol{x}_i \in R^n$, $i = 1,..,l$, oznacza wektory trenujące, $\boldsymbol{y} \in R^l$, $y_i \in \{1, -1\}$, $\boldsymbol{e}$ jest wektorem wartości równych 1, $C > 0$ oznacza funkcję kosztu i wyraża parametr kary w wyznaczaniu błędu, a Q to dodatnio półokreślona macierz o wymiarach l x l, zbudowana zgodnie z tożsamością 5.35:

$$Q_{ij} \equiv y_i y_j K\left(\mathbf{x}_i, \mathbf{x}_j\right) \tag{5.35}$$

gdzie K oznacza funkcję jądra, określoną tożsamością 5.36.

$$K\left(\mathbf{x}_i, \mathbf{x}_j\right) \equiv \phi\left(\mathbf{x}_i\right)^T \phi\left(\mathbf{x}_j\right) \tag{5.36}$$

Na podstawie badań opisanych w rozdziale 5.2 wybrano liniowe jądro, dla którego funkcja K określona jest równaniem 5.37. W ten sposób, funkcja wyznaczająca decyzję o przynależności obiektu do klasy przyjęła postać 5.38.

$$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \mathbf{x}_i^T\mathbf{x}_j \tag{5.37}$$

$$g(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{l} y_i\alpha_i\mathbf{x}_i^T\mathbf{x} + b\right) \tag{5.38}$$

Zgodnie ze słownikiem gestów przedstawionym w rozdziale 5.5 w systemie zdefiniowano trzy klasy gestów statycznych. Dla dłoni lewej gesty są odbiciem zwiercia-

dlanym gestów wykonywanych prawą dłonią. Z każdą klasą gestu związany jest jeden klasyfikator. Aby możliwe było rozpoznawanie różnych gestów statycznych wykonywanych jednocześnie obiema rękoma, tworzone są osobne klasyfikatory dla lewej i prawej dłoni.

W oparciu o wyniki badań opisanych w rozdziale 5.2 przyjęto jądro liniowe. Parametry C i γ klasyfikatorów przyjęły wartości odpowiednio $2^{1,25}$ i 2^{-15} dla dłoni lewej i $C = 2^2$, $\gamma = 2^{-8,75}$ dla dłoni prawej. Pozostałe parametry były identyczne dla obu klasyfikatorów. Wartości określono w procesie analizy wyników klasyfikacji otrzymywanych dla różnych wartości parametrów klasyfikatora i tych samych wektorów parametrów wejściowych.

W procesie trenowania klasyfikatorów każda klasa gestu reprezentowana jest przez 90 histogramów PGH. Zgodnie z opisem w rozdziale 5.2, w ramach tego zbioru można wyróżnić trzy podzbiory 30 histogramów reprezentujących kształty dłoni uzyskane dla trzech różnych trajektorii ruchu. Trenowanie klasyfikatorów z wykorzystaniem biblioteki LIBSVM realizowane jest metodą „jeden przeciwko wszystkim". Zgodnie z tą metodą wektory opisujące klasy gestów dzielone są na dwa zbiory. Do jednego zbioru należą wektory reprezentujące klasę, dla której dany klasyfikator jest trenowany, do drugiego zaś – wszystkie pozostałe wektory.

Algorytmy maszyny wektorów nośnych, udostępnianej przez bibliotekę LIBSVM, w stosunku do klasycznej metody są rozszerzone o wartości prawdopodobieństw przynależności rozpoznawanego obiektu do danej klasy. W systemie wykorzystano tę właściwość, określając próg podjęcia decyzji o przynależności kształtu dłoni do klasy gestu jako 0,5. Za odpowiedź całego systemu klasyfikacji przyjmuje się pozytywną odpowiedź klasyfikatora, który zwraca maksymalną wartość prawdopodobieństwa. Jeśli maksymalna wartość prawdopodobieństwa jest mniejsza niż 0,5, to system zwraca klasę gestu odpowiadającą płaskiej dłoni, czyli brakowi akcji.

Sformułowane w niniejszym rozdziale założenia systemu zostały poddane walidacji i testom. Badania systemu przedstawiono w rozdziale 6.

6 Badanie opracowanego systemu

6.1 Wpływ wizualizacji parametrów i ergonomii na wyniki procesu miksowania

Badania wpływu wizualizacji parametrów i ergonomii na wyniki procesu miksowania zostały podzielone na dwa etapy. W ramach pierwszego etapu wykonane zostały procesy miksowania dźwięku z wykorzystaniem opracowanego interfejsu oraz wybranego systemu produkcji muzycznej. Badania zostały wykonane dla różnych wariantów obsługi opracowanego interfejsu i systemu produkcji muzycznej, w taki sposób, aby możliwa była wiarygodna ocena wpływu wizualizacji parametrów sygnałów fonicznych na decyzje podejmowane w trakcie miksowania, warunkujące walory estetyczne otrzymywanych zgrań. Drugim aspektem, z myślą o którym opracowano eksperyment, było porównanie ergonomii obsługi procesów miksowania za pomocą gestów z obsługą systemu miksowania za pomocą myszy i klawiatury. Wyniki porównania przedstawiono również w jednej z publikacji autora niniejszej rozprawy [82]. Otrzymane próbki dźwiękowe zostały poddane ocenie w testach subiektywnych stanowiących drugi etap badań.

Sformułowane zostały dwie przedstawione poniżej hipotezy, do których autor rozprawy odnosi się w dalszej części niniejszego rozdziału.

Hipoteza 1

Wizualizacja parametrów sygnałów fonicznych wpływa negatywnie na walory estetyczne otrzymywanych zgrań.

Hipoteza 2

Miksowanie za pomocą gestów rąk prowadzi do uzyskania zgrań o wyższych walorach estetycznych niż miksowanie przy użyciu myszy i klawiatury.

6.1.1 Metodyka miksowania sygnałów

W pierwszej części badań udział wzięło 10 realizatorów nagrań zajmujących się w ramach swoich czynności zawodowych miksowaniem sygnałów fonicznych. Zadaniem każdego z realizatorów było zmiksowanie dostarczonych przez autora systemu miksowania za pomocą gestów ośmiu ścieżek wchodzących w skład tego samego nagrania. Ścieżki znacząco różniły się od siebie pod względem cech muzycznych i sygnałowych. Każda ze ścieżek zawierała zarejestrowaną partię innego instrumentu lub grupy instrumentów, kolejno: bębna basowego, werbla, talerzy (*hi-hat*), gitary basowej, fortepianu, gitary prowadzącej, gitary rytmicznej i orkiestry symfonicznej. Partie gitar i fortepianu zostały zarejestrowane z wykorzystaniem instrumentów przez autora rozprawy. Partia orkiestry symfonicznej została stworzona w oparciu o wtyczkę programową Steinberg Halion Symphonic Orchestra, dokonującą zaawansowanego przetwarzania dźwięków instrumentów akustycznych zarejestrowanych w sali koncertowej. Partie pozostałych instrumentów wykreowano z wykorzystaniem instrumentów wirtualnych. Utwór został stworzony przez autora rozprawy. Każdy z realizatorów miksował dostarczone ścieżki po raz pierwszy. Istotą było najpierw wypracowanie przez każdego realizatora indywidualnej koncepcji względem finalnych właściwości miksu, a następnie podjęcie próby urzeczywistnienia tej koncepcji, mając na względzie uzyskanie za każdym razem identycznego zgrania. Realizatorzy byli proszeni o przyjęcie niezmiennej metodyki postępowania dla wszystkich, przedstawionych poniżej wariantów miksowania.

W celu porównania ergonomii opracowanego interfejsu z ergonomią systemu DAW oraz sprawdzenia wpływu odzwierciedlania parametrów za pomocą informacji wizualnej na walory estetyczne zgrań, rozważyć należy pięć sposobów miksowania sygnałów fonicznych, wyszczególnionych w tabeli 6.1.

Tabela 6.1 Zbadane sposoby miksowania sygnałów fonicznych

Symbol sposobu miksowania	Opis sposobu miksowania
①	miksowanie za pomocą gestów z wykorzystaniem opracowanego systemu bez odzwierciedlenia zmian parametrów dźwięku w postaci informacji wizualnej
②	miksowanie za pomocą gestów z wykorzystaniem opracowanego systemu, z odzwierciedleniem zmian parametrów dźwięku w postaci informacji wizualnej
③	miksowanie z wykorzystaniem opracowanego systemu obsługiwanego za pomocą myszy i klawiatury, bez odzwierciedlenia zmian parametrów dźwięku w postaci informacji wizualnej
④	miksowanie z wykorzystaniem opracowanego systemu obsługiwanego za pomocą myszy i klawiatury, z odzwierciedleniem zmian parametrów dźwięku w postaci informacji wizualnej
⑤	miksowanie z wykorzystaniem systemu produkcji muzycznej, obsługiwanego zgodnie z wypracowanym przez eksperta indywidualnym sposobem pracy, za pomocą myszy, klawiatury oraz kontrolera MIDI

W sposobie ⑤ operacje miksowania, które mógł wykonać ekspert ograniczone były do zbioru operacji dostępnych w przypadku miksowania za pomocą gestów. Uzasadnienie przeprowadzenia badań w oparciu o pięć przedstawionych powyżej sposobów miksowania zawarto w dalszej części rozprawy w formie tabelarycznej (tabela 6.2) w rozdziale 6.1.3.

Zadaniem ekspertów było wykonanie możliwie najbardziej podobnych zgrań sygnałów fonicznych (w obrębie zgrań danego eksperta) w każdy z wyżej wymienionych sposobów. Otrzymane zgrania zamieszczono na płycie DVD, która stanowi dodatek I do rozprawy. Kolejność sposobów dla każdego z ekspertów była inna. Miało to na celu wyeliminowanie efektu nauki. Efekt ten mógłby prowadzić do otrzymania wyników korelujących z kolejnością sposobów miksowania, przez co niemożliwe byłoby wiarygodne wnioskowanie o ergonomii interfejsu bądź związku pomiędzy wizualizacją informacji a podejmowaniem decyzji wpływających na walory estetyczne zgrań.

Po zmiksowaniu sygnałów za pomocą wszystkich pięciu sposobów, każdy z ekspertów proszony był o wypełnienie ankiety badającej różne aspekty opracowanego interfejsu miksowania. W ramach badania ankietowego oceniane były m.in. dokładność, wygoda obsługi, intuicyjność. Eksperci byli również proszeni o dokonanie oceny

subiektywnej wykonanych przez siebie zgrań. Wyniki oceny wraz z analizą odpowiedzi udzielonych w ankietach zostały przedstawione w rozdziale 6.1.4. Wzór ankiety zamieszczono w dodatkach do rozprawy w sekcji H. Na płycie DVD (dodatek I) zamieszczono zeskanowane ankiety wypełnione przez realizatorów.

6.1.2 Warunki eksperymentu

Zarówno miksowanie sygnałów fonicznych, jak i testy subiektywne, mające na celu ocenę otrzymanych zgrań, odbyły się w jednakowych warunkach. Pomieszczeniem, w którym przeprowadzono badania była sala seminaryjna Katedry Systemów Multimedialnych. Pomieszczenie to wybrano ze względu na obecność projektora multimedialnego i ekranu, możliwość zaciemnienia poprzez opuszczenie rolet oraz dobre warunki akustyczne. Odsłuch stanowiły monitory studyjne Yamaha MSP5, umieszczone na statywach Ultimate Support MS-45B2. Odległość pomiędzy monitorami wynosiła 1,85 m. Osoba miksująca sygnały foniczne lub słuchacz w testach subiektywnych usytuowani byli w punkcie optymalnego odsłuchu, tj. głowa zlokalizowana była w wierzchołku trójkąta równobocznego, którego pozostałe dwa wierzchołki wyznaczały miejsca lokalizacji monitorów odsłuchowych. Zadbano o to, aby głowa użytkownika systemu znajdowała się na wysokości głośników wysokotonowych. Zapewnienie takiej pozycji odsłuchowej było istotne, biorąc pod uwagę zmianę sposobu lokalizowania dźwięku i postrzegania jego przestrzenności przy zmianie odległości pomiędzy lewym i prawym uchem a źródłem dźwięku [72] [113]. Z tego powodu, dla sposobów miksowania ① i ②, w których praca odbywa się w pozycji stojącej, statywy z monitorami odsłuchowymi umieszczone zostały na blatach stołów. Miksowanie zgodnie ze sposobami ③ – ⑤ oraz indywidualna ocena zgrań przez ekspertów, jak i ocena w testach subiektywnych z udziałem niezależnych słuchaczy, odbywały się w pozycji siedzącej.

6.1.3 Metodyka testów subiektywnych

Po zmiksowaniu dostarczonych sygnałów fonicznych zgodnie z pięcioma sposobami realizatorzy proszeni byli o ocenę subiektywną uzyskanych pięciu zgrań. Ocena ta wykonana została zgodnie z metodyką testu szeregowania rang [152] z użyciem skali pięciostopniowej (1 – 5). Wyniki wraz z analizą przedstawiono w punkcie 6.1.4. W celu sprawdzenia stopnia przypadkowości w ocenach nadanych zgraniom, przeprowadzono

dodatkowo testy subiektywne z udziałem większego grona osób, nie będących realizatorami. Testy te zostały oparte na metodzie porównań parami z ocenami w skali dwustopniowej (lepszy – gorszy). Zgodnie z metodyką porównań parami zastosowano dwie serie testowe, aby sprawdzić wiarygodność odpowiedzi osób, które wzięły udział w badaniu. Sposób wyboru stabilnych odpowiedzi przedstawiono w rozdziale 6.1.5. Pary utworzono z piętnastosekundowych fragmentów zgrań wykonanych przez tego samego eksperta, według różnych sposobów miksowania. Udział w testach wzięli pracownicy Katedry Systemów Multimedialnych, prowadzącej specjalność Inżynieria Dźwięku i Obrazu.

Z pięciu zgrań, będących efektem pracy jednego realizatora nagrań, możliwe jest utworzenie dziesięciu kombinacji par. Informacje, jakich dostarcza porównanie poszczególnych par zawarto w tabeli 6.2. Zastosowane w tabeli symbole zgrań ① – ⑤ są równoznaczne z symbolami reprezentującymi poszczególne sposoby miksowania, wyjaśnionymi w tabeli 6.1.

Tabela 6.2 Zestawienie par zgrań uzyskanych za pomocą różnych sposobów miksowania i informacji, jakich dostarczają

Para zgrań	Informacja, jakiej dostarcza porównanie dwóch zgrań w parze
① ∧ ②	Sprawdzenie wpływu wizualizacji parametrów sygnałów fonicznych na walory estetyczne zgrań przy obsłudze za pomocą gestów
① ∧ ③	Sprawdzenie ergonomii / dokładności interfejsu obsługiwanego za pomocą gestów
① ∧ ④	Para kontrolna (ze względu na silne zróżnicowanie warunków eksperymentu, w jakich powstały próbki, para ta nie może służyć w oderwaniu od innych par, jako podstawa wnioskowania)
① ∧ ⑤	Analizowana wraz z parami ② ∧ ⑤, ③ ∧ ⑤ i ④ ∧ ⑤, dostarcza informacji czy kluczowe znaczenie miał wpływ wizualizacji parametrów sygnałów fonicznych na walory estetyczne zgrań czy sposób obsługi
② ∧ ③	Analizowana z innymi parami dostarcza informacji czy większy wpływ na wyniki miksowania ma sposób obsługi czy obecność informacji wizualnej
② ∧ ④	Sprawdzenie ergonomii / dokładności interfejsu obsługiwanego za pomocą gestów
② ∧ ⑤	Porównanie ergonomii interfejsu obsługiwanego za pomocą gestów z ergonomią systemu DAW

③ ∧ ④	Sprawdzenie wpływu wizualizacji parametrów sygnałów fonicznych na walory estetyczne zgrań przy obsłudze opracowanego interfejsu za pomocą myszy i klawiatury
③ ∧ ⑤	Para kontrolna (ze względu na silne zróżnicowanie warunków eksperymentu (systemów), w jakich powstały próbki, para ta nie może służyć w oderwaniu od innych par, jako podstawa wnioskowania)
④ ∧ ⑤	Para kontrolna (ze względu na silne zróżnicowanie warunków eksperymentu (systemów), w jakich powstały próbki, para ta nie może służyć w oderwaniu od innych par, jako podstawa wnioskowania)

Informacji na temat potwierdzenia lub odrzucenia postawionej w ramach tego rozdziału hipotezy nr 1 dostarcza wynik porównań sposobów ① ∧ ② oraz ③ ∧ ④. Przewagę wybranego sposobu nad drugim z pary oznaczono symbolicznie poprzez znak „>" pomiędzy numerami sposobów. O potwierdzeniu hipotezy nr 1 mógłby świadczyć wynik porównania par, zgodny z relacją 6.1.

$$① > ② \wedge ③ > ④ \tag{6.1}$$

Warto zauważyć, że przy obsłudze opracowanego interfejsu wyłącznie za pomocą myszy i klawiatury, użytkownik musi skupiać wzrok na ekranie monitora, aby możliwy był wybór opcji systemu. Warunek ten istnieje niezależnie od włączenia czy wyłączenia opcji odzwierciedlania zmian parametrów w postaci informacji wizualnej. Z kolei, w przypadku obsługi opracowanego interfejsu za pomocą gestów, możliwe jest zamknięcie oczu i wykonywanie operacji miksowania bez angażowania zmysłu wzroku również przy włączonej opcji wyświetlania informacji odzwierciedlających modyfikacje parametrów. O skorzystaniu z tej możliwości przez użytkowników mogłaby świadczyć mniejsza różnica w przewadze sposobu ① nad sposobem ② niż w przypadku wyniku porównania sposobów ③ i ④. Uzupełnienie relacji 6.1 o porównanie pary ③ i ⑤, zgodnie z relacją 6.2, pozwalałoby dodatkowo na porównanie opracowanego interfejsu z zastosowanym systemem produkcji muzycznej w kontekście specyfiki wielomodalnej percepcji. Przewaga sposobu ⑤ nad sposobem ③, przy pozostałych parach będących w relacji 6.1, mogłaby świadczyć o tym, że istnieją inne czynniki niż angażowanie zmysłu wzroku, niewystępujące w przypadku opracowanego interfejsu, a mające duży

wpływ na wyniki miksowania podczas pracy z wybranym systemem produkcji muzycznej.

$$① > ② \wedge ③ > ④ \wedge ③ > ⑤ \tag{6.2}$$

O niewystarczającej ergonomii interfejsu myszy i klawiatury w procesie miksowania dźwięku można by wnioskować na podstawie relacji 6.3.

$$① > ③ \wedge ② > ④ \tag{6.3}$$

Dodanie pary ② i ⑤ i wynik porównania zgodny z relacją 6.4 pozwala wykluczyć obecność innych czynników w trakcie miksowania w oparciu o wybrany system produkcji muzycznej, które mogłyby mieć większe znaczenie dla wyników miksowania niż rodzaj przyjętego sposobu sterowania.

$$① > ③ \wedge ② > ④ \wedge ② > ⑤ \tag{6.4}$$

Sekwencja, która by jednoznacznie potwierdzała przewagę miksowania bez angażowania zmysłu wzroku nad miksowaniem wspieranym interfejsem graficznym, a dodatkowo świadczyła o większej ergonomii interfejsu sterowanego za pomocą gestów w porównaniu z systemem obsługiwanym przy użyciu myszy i klawiatury, przyjmuje postać 6.5.

$$① > ② \wedge ① > ③ \wedge ① > ④ \wedge ① > ⑤ \wedge \tag{6.5}$$

$$② > ④ \wedge ② > ⑤ \ \wedge ③ > ④ \wedge ③ > ⑤ \wedge ⑤ > ④$$

Zauważyć można, że w relacji 6.5 nie uwzględniono par ② i ③ oraz ④ i ⑤. Na podstawie wyniku porównania sposobów uwzględnionych w tych parach nie można potwierdzić ani sfalsyfikować postawionych hipotez. Relacja ② > ③ w zależności od wyniku pozostałych relacji mogłaby świadczyć jedynie o tym, że ergonomia interfejsu ma większy wpływ na uzyskiwane efekty miksowania niż angażowanie zmysłu wzroku. Przeciwnie, relacja ③ > ② mogłaby oznaczać, że angażowanie zmysłu wzroku poprzez wpływ na percepcję dźwięku ma silniejszy związek z uzyskiwanymi efektami miksowania niż użycie mniej ergonomicznego interfejsu myszy i klawiatury. Dla relacji ③ > ②

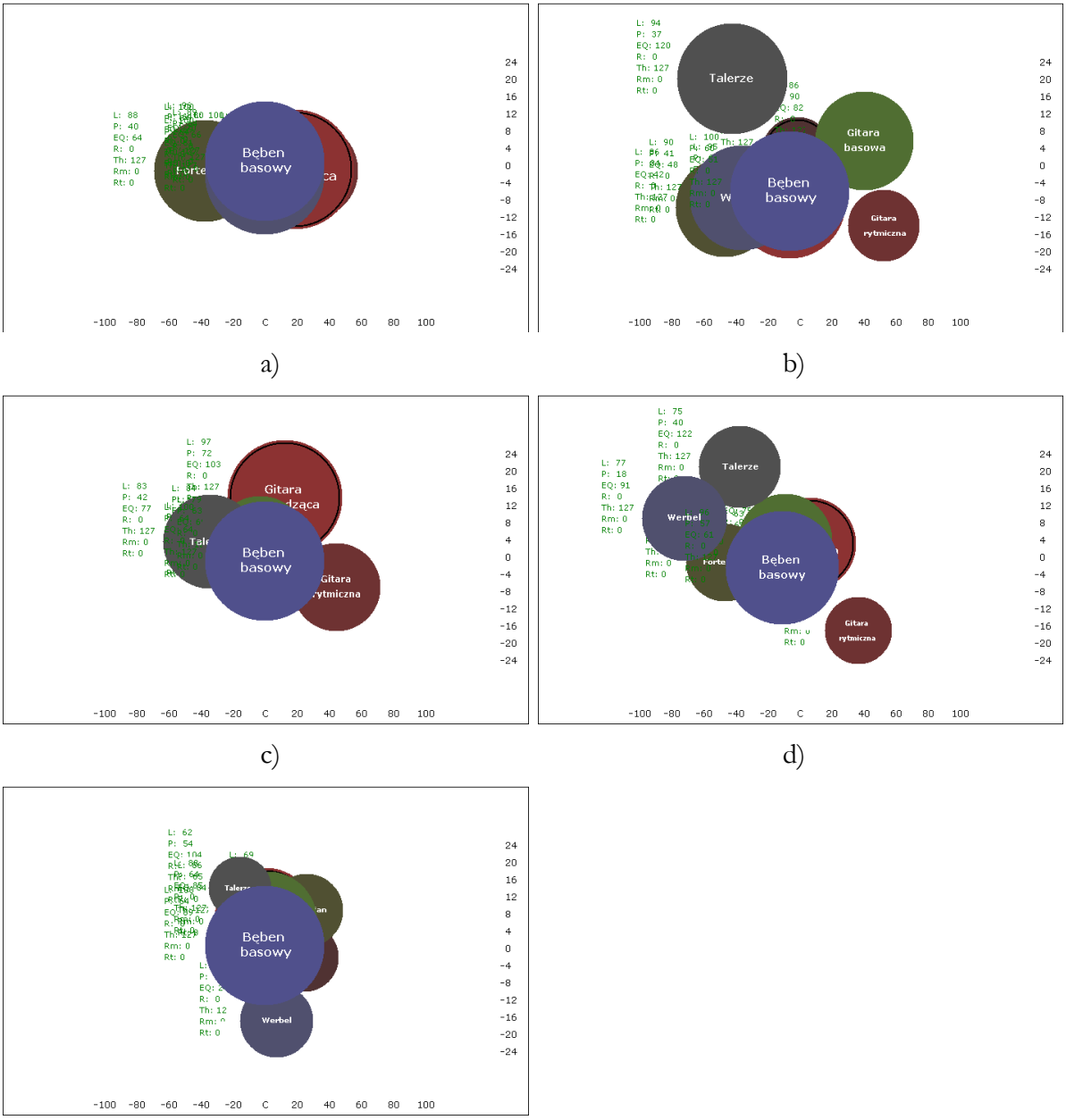
dodatkowo wynik porównania pary ① i ③ informuje o tym czy na wyniki faktycznie negatywny wpływ miało angażowanie zmysłu wzroku (① > ③), czy niewystarczająca dokładność odzwierciedlenia ruchu rąk w trakcie wykonywania gestu na zmianę wartości parametru (③ > ①). Relacja odwrotna do relacji ⑤ > ④ mogłaby sugerować, że na lepsze wyniki miksowania z wykorzystaniem opracowanego interfejsu miały wpływ dodatkowe czynniki nie podlegające badaniu.

Dla 10 realizatorów, którzy wzięli udział w badaniach możliwe jest utworzenie 10 zestawów testowych składających się z dwóch serii 10 par. Przyjmując, że przerwy pomiędzy 15-sekundowymi próbkami w parze wynoszą dwie sekundy, pomiędzy parami – 5 sekund, pomiędzy dwoma seriami przypadającymi na każdy zestaw – 3 minuty, a między zestawami – 4 minuty, łączny czas trwania badania dla jednej osoby wynosiłby około 152 minuty. Z tego względu do testów subiektywnych wybrano zgrania dwóch realizatorów (realizatorzy nr 3 i 7), którzy nadali skrajne oceny zgraniom otrzymanym w rezultacie miksowania za pomocą gestów przy braku wizualizacji parametrów sygnałów fonicznych. Ponadto zgrania tych realizatorów, w obrębie ich własnych dokonań, znacząco różniły się od siebie, co dodatkowo przemawiało za wyborem ich dokonań w kontekście uzyskania stabilnych odpowiedzi słuchaczy. W badaniu wzięło udział 29 osób. Każda z nich oceniała dwa zestawy utworzone dla zgrań obu ekspertów. Przerwa między badaniami z wykorzystaniem próbek danego zestawu wynosiła, dla każdej z osób, co najwyżej trzy dni. Pojedyncze badanie trwało około 15 minut. Jak już wcześniej napisano, warunki testów odsłuchowych były identyczne z warunkami, w jakich zaproszeni realizatorzy miksowali udostępniony materiał muzyczny.

6.1.4 Analiza wyników badań z udziałem realizatorów nagrań

Niezależnie od pracy w trybie ograniczonego bądź pełnego interfejsu graficznego dla każdego z otrzymanych zgrań utworzona została wizualizacja odpowiadająca widokowi ekranu w trakcie miksowania zgodnie z drugim z wymienionych trybów, tj. lokalizacja źródła na osi poziomej odpowiadała lokalizacji w bazie stereofonicznej, a na osi pionowej – wzmocnieniu korektora. Wielkość źródła reprezentowała poziom sygnału. Dodatkowo każdemu ze źródeł sygnału przypisany został niezmienny kolor. Na potrzeby analizy statystycznej zarejestrowano wartości wszystkich parametrów. W sekcji C dodatków do rozprawy zamieszczono wizualizacje wszystkich zgrań wraz z dokład-

nymi wartościami parametrów. Na rys. 6.2 przedstawiono przykładowe wizualizacje zgrań realizatora nr 3. W tabeli 6.3 zawarto wartości wszystkich parametrów dla tych zgrań. Wyniki miksowania uzyskane dla realizatora nr 3 wybrano jako przykład, ponieważ otrzymane przez niego zgrania stanowiły wraz ze zgraniami realizatora nr 7 próbki podlegające ocenie w testach subiektywnych.



a) b) c) d) e)

Rys. 6.2 Wizualizacje zgrań realizatora nr 3: (a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

Tabela 6.3 Wartości kontrolerów parametrów zgrań realizatora nr 3

(a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

a)

	Poziom	Panorama	Wzmocnienie korektora	Próg kompresji	Stopień kompresji	Miks pogłosu	Czas pogłosu
Bęben basowy	100	64	69	127	0	0	0
Werbel	100	64	61	127	0	0	0
Talerze (*Hi-Hat*)	89	64	66	127	0	0	0
Gitara basowa	96	64	70	127	0	0	0
Fortepian	88	40	64	127	0	0	0
Gitara prowadząca	100	76	64	127	0	12	29
Gitara rytmiczna	85	82	64	127	0	0	0
Orkiestra symfoniczna	80	64	64	127	0	29	25

b)

	Poziom	Panorama	Wzmocnienie korektora	Próg kompresji	Stopień kompresji	Miks pogłosu	Czas pogłosu
Bęben basowy	100	60	51	127	0	0	0
Werbel	90	41	48	127	0	0	0
Talerze (*Hi-Hat*)	94	37	120	127	0	0	0
Gitara basowa	86	90	82	127	0	0	0
Fortepian	86	34	42	127	0	0	0
Gitara prowadząca	95	60	45	127	0	25	47
Gitara rytmiczna	68	98	31	127	0	0	0
Orkiestra symfoniczna	68	63	75	127	0	27	52

c)

	Poziom	Panorama	Wzmocnienie korektora	Próg kompresji	Stopień kompresji	Miks pogłosu	Czas pogłosu
Bęben basowy	100	64	64	74	22	0	0
Werbel	79	63	69	127	0	0	0
Talerze (*Hi-Hat*)	83	42	77	127	0	0	0
Gitara basowa	84	62	75	118	12	0	0
Fortepian	84	60	65	127	0	0	0
Gitara prowadząca	97	72	103	127	0	13	40
Gitara rytmiczna	79	93	49	127	0	0	0
Orkiestra symfoniczna	77	64	64	127	0	32	28

d)

	Poziom	Panorama	Wzmocnienie korektora	Próg kompresji	Stopień kompresji	Miks pogłosu	Czas pogłosu
Bęben basowy	96	57	61	50	68	0	0
Werbel	77	18	91	87	26	0	0
Talerze (*Hi-Hat*)	75	40	122	127	0	0	0
Gitara basowa	83	58	77	127	0	0	0
Fortepian	72	34	64	127	0	0	0
Gitara prowadząca	82	68	75	127	0	23	25
Gitara rytmiczna	65	88	23	127	0	0	0
Orkiestra symfoniczna	63	61	60	127	0	40	60

e)

	Poziom	Panorama	Wzmocnienie korektora	Próg kompresji	Stopień kompresji	Miks pogłosu	Czas pogłosu
Bęben basowy	100	64	69	127	0	0	0
Werbel	69	69	24	39	39	0	0
Talerze (*Hi-Hat*)	62	54	104	127	0	0	0
Gitara basowa	86	65	84	127	0	0	0
Fortepian	69	81	90	127	0	0	0
Gitara prowadząca	88	64	85	127	0	35	28
Gitara rytmiczna	61	73	58	127	0	0	0
Orkiestra symfoniczna	66	80	62	52	6	33	57

Ocena wpływu wizualizowania informacji i sposobu interakcji na rozkład wartości parametrów

W przypadku zgrań sześciu realizatorów (realizatorzy nr: 1, 2, 3, 5, 9, 10) zaobserwowano wyraźne różnice w lokalizacji źródeł dźwięku w zależności od trybu graficznego interfejsu. Na podstawie wizualizacji zgrań pięciu realizatorów w tej grupie zauważono, że miksowanie wsparte pełną informacją wizualną skutkowało większym rozsunięciem źródeł zarówno w osi poziomej, jak i pionowej. Odpowiadało to szerszej panoramie i bardziej zdecydowanemu użyciu korektora częstotliwości. Prawidłowość ta istniała niezależnie od wyboru sposobu interakcji (gesty / mysz). Taki efekt można uznać za będący przeciwny z przekonaniem, zgodnie z którym wsparcie wizualne powinno skutkować łatwiejszą i przez to szybszą percepcją słuchową zmiany parametru. Okazało się jednak, że w sytuacji braku wizualizacji źródeł i braku wartości parame-

trów realizatorzy zdawali się poświęcać więcej uwagi balansowi dźwięku. W istocie pozorne zbalansowanie źródeł, o którym można by wnioskować na podstawie jedynie oceny wizualnej ich rozmieszczenia i wielkości, okazało się nieodpowiadającym rzeczywistemu zbalansowaniu dźwięku. Różnice w pozostałych parametrach uniemożliwiły jednak powiązanie niniejszego rozstawienia źródeł na płaszczyźnie z wyższą wartością estetyczną otrzymanych zgrań.

Oprócz określenia zmienności panoramy i wzmocnienia korektora w oparciu o wizualizacje miksów, dokonano również statystycznej analizy rozkładu wartości pozostałych parametrów, tj. poziomu, progu i stopnia kompresji dynamiki oraz miksu i czasu pogłosu. Dodatkowo, aby możliwe było zbadanie istotności statystycznej rozkładu wartości parametrów panoramy *pan* i wzmocnienia korektora *eq*, utworzono dwie miary oceny stopnia zróżnicowania tych wartości w ramach danego miksu. Miary te zdefiniowano jako bezwzględną różnicę wartości środkowej zakresu wartości przyjmowanych przez kontroler MIDI i wartości parametru zgodnie ze wzorem 6.6.

$$\begin{aligned} pan_{diff} &= |pan - 64| \\ eq_{diff} &= |eq - 64| \end{aligned} \tag{6.6}$$

W pierwszym etapie analizy statystycznej, dla wymienionych powyżej pięciu parametrów i dwóch miar sprawdzono istotność różnic w rozkładach wartości dla wszystkich sygnałów fonicznych i wszystkich inżynierów jednocześnie. Celem testu było sprawdzenie w pierwszej kolejności czy istotne statystycznie różnice zachodzą niezależnie od cech osobniczych i niezależnie od rodzaju edytowanego źródła dźwięku. Ponieważ wartości parametrów były liczbami całkowitymi, a ich rozkład nie był rozkładem normalnym, analizę przeprowadzono z wykorzystaniem testu rang Friedmana [33]. Uzyskane wartości prawdopodobieństw testowych *p* dla czterech spośród pięciu parametrów i obu miar przyjęły wartości mniejsze od poziomu istotności wynoszącego 0,05 (Dodatek D: tabela D.1). Jedynie dla parametru *próg kompresji* uzyskana wartość wynosiła 0,664 > 0,05. Stwierdzono zatem, że nie ma podstaw do odrzucenia hipotezy zerowej stanowiącej o tym, że obrany sposób miksowania dźwięku wpływa na dobór wartości dla parametru *próg kompresji*. Dla pozostałych parametrów i miar przyjęto hipotezę alternatywną, zgodnie z którą, pomiędzy co najmniej dwoma sposobami miksowania zachodzą istotne statystycznie różnice, które determinują dobór wartości w

trakcie edycji. W celu sprawdzenia, które sposoby miksowania różniły się pomiędzy sobą w takim stopniu, aby otrzymać istotnie różne rozkłady wartości parametrów, wykonano serię testów *post hoc* par rangowanych znaków Wilcoxona [33]. Wyniki testów przedstawiono w tabelach D.2 – D.7 w dodatkach. Widoczne na wizualizacjach miksów różnice w panoramowaniu źródeł i użyciu korektora zostały potwierdzone w analizie statystycznej. Wartości prawdopodobieństw testowych dla par sposobów miksowania różniących się trybem interfejsu graficznego były mniejsze od poziomu istotności 0,05 niezależnie od sposobu interakcji (gesty / mysz). Identyczne wyniki otrzymano również dla parametru *poziom*. Dla par, w których jeden z porównywanych sposobów oparty był na gestach rąk istotne różnice zaobserwowano również w rozkładzie wartości parametru *miks pogłosu*. Otrzymane wyniki świadczą o związku pomiędzy obecnością informacji wizualnej odzwierciedlającej wartości parametrów a podejmowanymi w trakcie miksowania decyzjami warunkującymi ich rozkład. Jednocześnie, istotny wpływ wyboru myszy lub gestów jako interfejsu komunikacji z systemem zaobserwowano jedynie dla parametru *poziom*. Wynik taki można uznać za pożądany, biorąc pod uwagę niewysoką ocenę wygody obsługi za pomocą gestów, przedstawioną w dalszej części niniejszego rozdziału. Brak wygody obsługi nie przyczynił się do uzyskania globalnie istotnych statystycznie różnic w rozkładzie wartości parametrów.

Ocena stopnia zaangażowania zmysłu wzroku w procesie miksowania

Dziewięciu spośród 10 realizatorów nagrań, którzy wzięli udział w badaniach, odpowiedziało twierdząco na pytanie czy w przypadku któregoś z systemów i sposobów obsługi zmysł wzroku był angażowany w mniejszym stopniu (w porównaniu z innymi zbadanymi sposobami), tj. w większym stopniu można było skoncentrować się na dźwięku. Ośmiu z nich uznało, że mniejsze zaangażowanie zmysłu wzroku wystąpiło w przypadku opracowanego w ramach rozprawy systemu, obsługiwanego za pomocą gestów w trybie ograniczonego interfejsu graficznego. Dla sześciu osób w tej grupie, obsługa interfejsu za pomocą myszy i klawiatury zamiast gestów, nie przeszkodziła w ocenie systemu jako mniej angażującego zmysł wzroku. Dwie spośród tych osób uznały również system DAW za angażujący zmysł wzroku w mniejszym stopniu. Świadczyć to może o intensywnym wykorzystywaniu przez nie kontrolera MIDI i ograniczeniu operacji wykonywanych za pomocą myszy i klawiatury do niezbędnego minimum lub

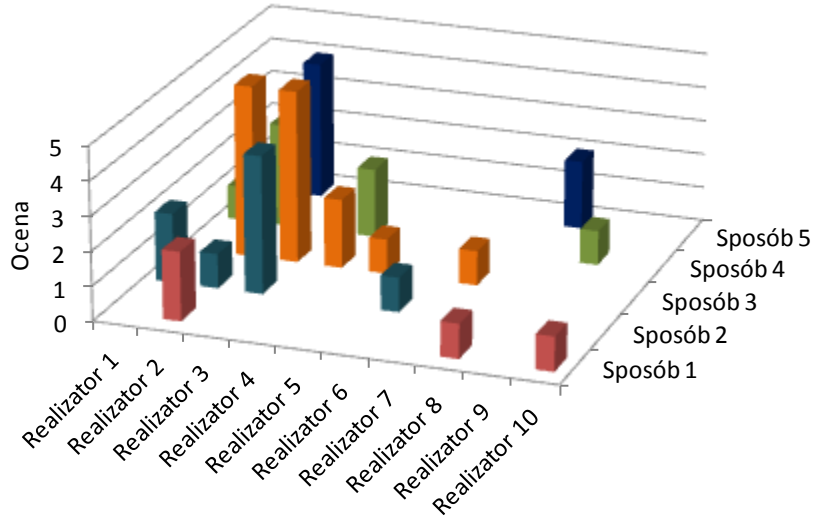
ocenie sposobu wizualizowania parametrów w systemie DAW jako mniej angażującego zmysł wzroku niż sposób przyjęty w opracowanym systemie w trybie pełnego interfejsu graficznego. Jeden z realizatorów uznał za mniej angażujące zmysł wzroku sposoby oparte na wykorzystaniu interakcji za pomocą gestów bez względu na obecność lub brak wizualizacji. Może to mieć związek z faktem, o którym wspomniano już wcześniej, tj. dzięki sposobowi, w jaki zaprojektowano system i słownik gestów, możliwe jest wybieranie większości parametrów i ich modyfikowanie przy zamkniętych oczach. Osoba ta uznała, że również w przypadku systemu DAW mogła w większym stopniu skoncentrować się na docierającym dźwięku. Dla jednego z realizatorów, opracowany system okazał się mniej angażujący zmysł wzroku jedynie w trybie obsługi za pomocą myszy i klawiatury.

Ocena stopnia powtarzalności wykonanych operacji

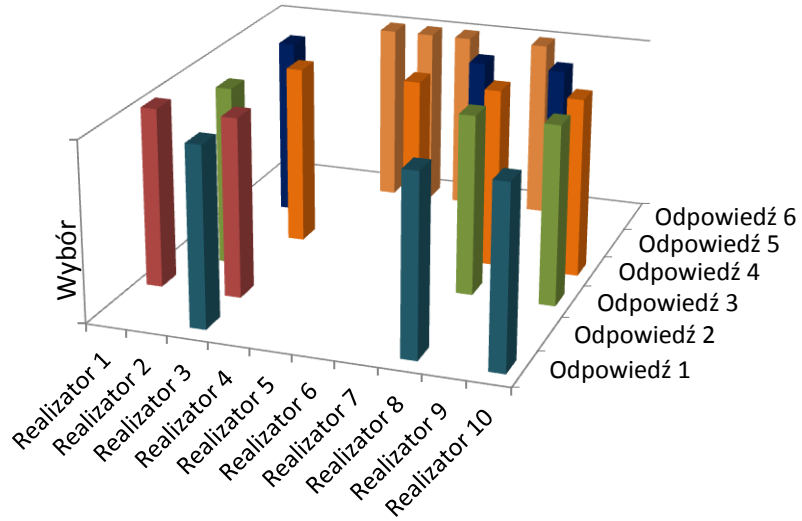
Przed dokonaniem oceny subiektywnej własnych zgrań każdy z realizatorów w ramach badania ankietowego był proszony o udzielenie odpowiedzi na pytanie czy był w stanie wykonać za każdym razem identyczny miks. Połowa spośród realizatorów odpowiedziała twierdząco (realizatorzy nr: 1, 2, 4, 6, 7). Analiza wizualizacji miksów i zarejestrowanych wartości wszystkich parametrów (Dodatek C: rys. i tabele 1, 2, 4, 6, 7) oraz analiza odsłuchowa pozwoliły jednak stwierdzić, że żadnemu z realizatorów nie udało się uzyskać na tyle podobnych zgrań, aby ich rozróżnienie stwarzało problemy. Fakt, że tylu realizatorów było przekonanych o uzyskaniu identycznych zgrań przed ich odsłuchaniem oraz zaobserwowane różnice w rozmieszczeniu źródeł na ekranie w zależności od trybu interfejsu graficznego mogą sugerować, iż sposób interakcji w procesie miksowania, jak i specyfika percepcji mogą w istocie wpływać na dobór wartości parametrów w trakcie miksowania. Fakt ten można uznać za dodatkowe uzasadnienie przyjętego sposobu realizacji systemu, uwzględniającego aspekty percepcyjne. Niezależnie od tego czy odpowiedź na postawione pytanie była pozytywna czy negatywna, żaden z realizatorów nie miał trudności z uszeregowaniem utworzonych zgrań pod względem walorów estetycznych.

Po odsłuchaniu wyników własnej pracy, realizatorzy byli proszeni o wskazanie tych zgrań, które ich zdaniem najbardziej różniły się od pozostałych. W ocenie tej realizatorzy mogli wskazać zarówno zgrania, które pod względem walorów estetycznych,

różniły się w sposób świadczący o ich przewadze nad pozostałymi, jak i niedoskonałościach. Wyniki wskazań, w odniesieniu do przydzielonych ocen, przedstawiono na rys. 6.3. Na rys. 6.4 przedstawiono rozkład odpowiedzi na pytanie o przyczyny otrzymania różnic pomiędzy zgraniami. Odpowiedzi zawarto w tabeli 6.4.



Rys. 6.3 Rozkład ocen dla zgrań uznanych za różniące się od pozostałych w obrębie zgrań każdego z realizatorów



Rys. 6.4 Rozkład odpowiedzi na pytanie o przyczyny różnic pomiędzy zgraniami

Tabela 6.4 Odpowiedzi realizatorów na pytanie o przyczyny otrzymania różnic pomiędzy zgraniami

Nr	Odpowiedzi na pytanie: "Co, Twoim zdaniem, było przyczyną otrzymania zgrania różniącego się od pozostałych?"
1	Niewystarczająca dokładność opracowanego systemu sterowanego za pomocą gestów
2	Obecność informacji wizualnej odzwierciedlającej zmiany wartości parametrów
3	Brak informacji wizualnej odzwierciedlającej wartości parametrów
4	Brak wygody obsługi opracowanego systemu za pomocą gestów
5	Inne
6	Trudno powiedzieć

Podczas, gdy dla większości realizatorów wybór zgrań różniących się od pozostałych i wskazanie słabych stron systemów jako przyczyn tych różnic odzwierciedlały zaniżone oceny wartości estetycznej, dla jednego realizatora (realizator nr 3) zależność ta była odwrócona. Realizator ten jako zgrania różniące się istotnie od pozostałych wytypował zgranie otrzymane w procesie miksowania za pomocą gestów w pełnym trybie graficznym oraz zgranie będące rezultatem miksowania za pomocą myszy i klawiatury w trybie ograniczonego interfejsu. Jako przyczynę różnic realizator wskazał niewystarczającą dokładność systemu, brak wygody obsługi oraz obecność wizualizacji. Wskazane aspekty dotyczyły systemu obsługiwanego za pomocą gestów. Pomimo tej oceny, oba zgrania zostały ocenione jako najlepsze pod względem wartości estetycznych. Na podstawie takiego wyniku stwierdzić można, że wyższa ergonomia interfejsu miksowania nie jest gwarancją uzyskania lepszych zgrań. Wśród innych przyczyn otrzymania zgrań różniących się od siebie, realizatorzy wymieniali zmęczenie wynikające z niewystarczającej ergonomii (realizator nr 2) oraz kolejność wykonywania miksów według poszczególnych sposobów (realizatorzy nr 7 i 9). Cztery osoby nie były w stanie podać przyczyny różnic (zaznaczenie w ankiecie odpowiedzi „trudno powiedzieć").

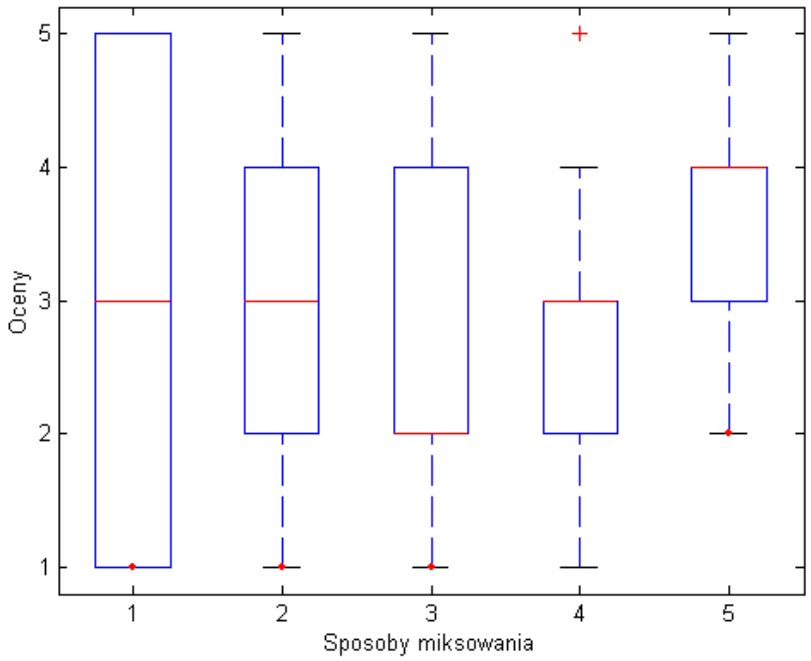
Ocena walorów estetycznych zgrań

Oceny walorów estetycznych przydzielone przez realizatorów własnym zgraniom przedstawiono w tabeli 6.5. Aby możliwa była ocena trendu upodobań zgrań według

poszczególnych sposobów miksowania, na wykresie pudełkowym (rys. 6.5) przedstawiono rozkład ocen zgrań (wykres „ramka-wąsy"). Czerwona linia wyznacza wartość mediany. Krawędzie pudełka – dolna i górna – wyznaczają odpowiednio – dolny i górny kwartyl. Krańce „wąsów" określają wartości: minimalną i maksymalną. Czerwonym krzyżykiem oznaczono obserwacje odstające. Identyczny sposób wizualizacji wyników w postaci wykresów pudełkowych zastosowano również w dalszej części rozprawy.

Tabela 6.5 Oceny walorów estetycznych przydzielone przez realizatorów własnym zgraniom (1 – najgorsze, 5 – najlepsze)

Ekspert	Gesty / tryb ograniczony	Gesty / tryb pełny	Mysz / tryb ograniczony	Mysz / tryb pełny	DAW
Realizator 1	5	2	4	1	3
Realizator 2	2	1	5	3	4
Realizator 3	1	4	5	3	2
Realizator 4	3	5	2	2	4
Realizator 5	5	2	1	3	4
Realizator 6	5	1	2	4	3
Realizator 7	5	3	1	2	4
Realizator 8	1	3	4	5	2
Realizator 9	3	4	2	1	5
Realizator 10	1	5	2	3	4



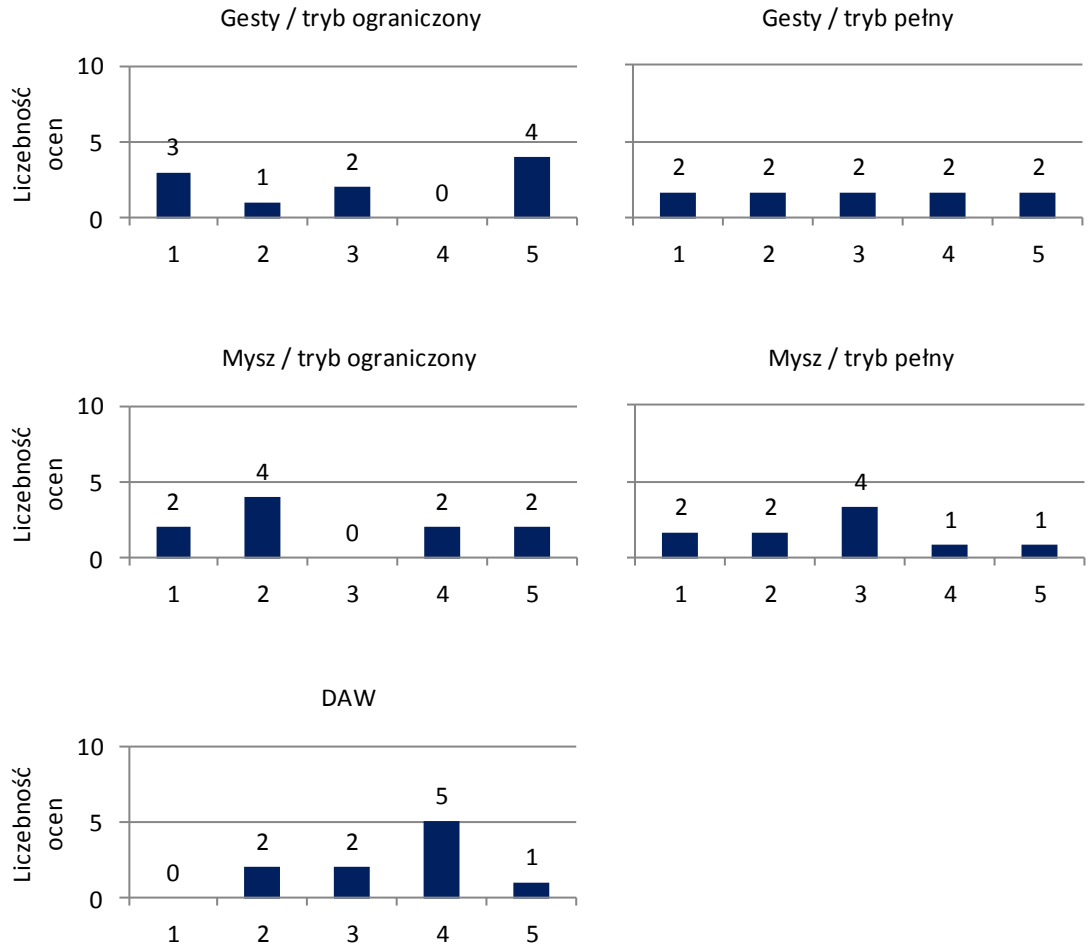
Rys. 6.5 Wykres pudełkowy ocen walorów estetycznych zgrań wszystkich realizatorów dla różnych sposobów miksowania

Sprawdzono, że przydzielone zgraniom oceny nie korelowały z kolejnością wykonywania miksów według poszczególnych sposobów. Nie zaobserwowano również korelacji pomiędzy ocenami przyznanymi zgraniom otrzymanym w procesie miksowania z wykorzystaniem bezpośrednio systemu Cubase a znajomością jego obsługi. Otrzymany rozkład typowań zgrań został poddany analizie statystycznej z wykorzystaniem testu rang Friedmana. Wyniki analizy przedstawiono w tabeli 6.6. W nagłówkach kolumn zawarto odpowiednio (patrz: Słownik pojęć): sumę kwadratów odchyleń pomiędzy grupami (SS Efekt), liczbę stopni swobody pomiędzy grupami (df Efekt), efekt średniokwadratowy (MS Efekt), sumę kwadratów odchyleń wewnątrz grup (SS Błąd), liczbę stopni swobody wewnątrz grup (df Błąd), błąd średniokwadratowy (MS Błąd), wartość testu chi-kwadrat (χ^2) oraz prawdopodobieństwo testowe (p).

Tabela 6.6 Zestawienie wartości testu rang Friedmana dla ocen przydzielonych przez realizatorów ich własnym zgraniom

	SS Efekt	df Efekt	MS Efekt	SS Błąd	df Błąd	MS Błąd	χ^2	p
ocena zgrania	4,45	4	1,1125	95,05	36	2,64028	1,79	0,7745

Uzyskanie wartości prawdopodobieństwa testowego p większej niż przyjęty poziom istotności równy 0,05 zadecydowało o przyjęciu hipotezy zerowej o braku różnic pomiędzy ocenami zgrań otrzymanych dla różnych sposobów miksowania. Zaobserwowano interesującą właściwość, polegającą na tym, że zgrania otrzymane w procesie miksowania za pomocą gestów w trybie ograniczonego interfejsu graficznego charakteryzowały się żywszym brzmieniem niż zgrania będące wynikiem miksowania z wykorzystaniem bezpośrednio środowiska DAW. Czterech realizatorów przydzieliło maksymalną notę zgraniom dla pierwszego z wyżej wymienionych sposobów (rys. 6.6). Zgranie z systemu DAW tylko w jednym przypadku otrzymało maksymalną ocenę (realizator nr 9) (rys. 6.6). Jednakże, żywe brzmienie zostało przez trzech realizatorów ocenione jako zbyt natarczywe, co skutkowało przydzieleniem przez nich oceny minimalnej równej 1 (rys. 6.6). **Uzyskane wyniki analizy statystycznej ocen walorów estetycznych zgrań stanowią część dowodu pierwszej tezy rozprawy.** Zgodnie z otrzymanymi wynikami zgrania otrzymane w procesie miksowania za pomocą gestów rąk nie różniły się w sposób istotny od zgrań będących wynikiem miksowania w środowisku DAW. Wynika z tego, że miksowanie za pomocą gestów rąk jest w istocie możliwe i efektywne. Ponadto szczególne znaczenie ma fakt, że zgrania otrzymane w drodze miksowania za pomocą gestów w trybie ograniczonego interfejsu wielokrotnie typowane były jako najlepsze.



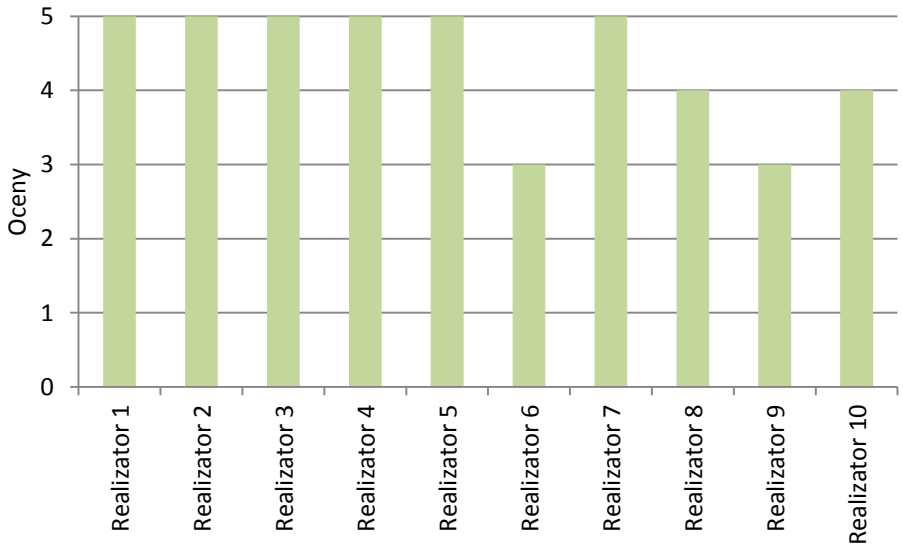
Rys. 6.6 Histogramy ocen walorów estetycznych zgrań dla każdego ze sposobów obsługi procesów miksowania

Ocena intuicyjności słownika gestów

Realizatorzy wysoko ocenili intuicyjność gestów (rys. 6.7). Sześciu realizatorów przydzieliło parametrowi intuicyjności maksymalną notę. Dwóch realizatorów oceniło intuicyjność na „3", proponując gest zamknięcia dłoni zamiast gestu „V" (realizator nr 6) oraz wskazując na potrzebę wprowadzenia większej liczby gestów statycznych celem ograniczenia zbioru gestów dynamicznych (realizator nr 9). Należy jednak zauważyć, że zaniżona ocena intuicyjności nie wpłynęła negatywnie na indywidualną ocenę walorów estetycznych zgrań tych realizatorów. Przeciwnie, zarówno realizator nr 6, jak i 9 ocenili zgranie otrzymane w procesie miksowania za pomocą gestów rąk w trybie ograniczonego interfejsu graficznego jako lepsze od zgrań uzyskanych w drodze miksowania za pomocą myszy komputerowej i klawiatury (tabela 6.5). Dwóch realizatorów oceniło intuicyjność na „4". Jeden z nich (realizator nr 10) wskazał na zbyt małą różnicę mię-

dzy gestami, z punktu widzenia użytkownika. Drugi (realizator nr 8), uznał, że konieczność przywracania w niektórych sytuacjach płaskiego kształtu dłoni przed wykonaniem nowego gestu wpływa na zaniżenie intuicyjności.

Należy zauważyć, że sposób miksowania dźwięku za pomocą gestów stanowił dla wszystkich realizatorów nowy sposób realizacji zgrania. Biorąc pod uwagę fakt, że realizatorzy, którzy wzięli udział w badaniach na co dzień wykorzystują odmienny sposób pracy z materiałem fonicznym, otrzymane oceny intuicyjności można uznać za wysokie. **Uzyskanie wysokich ocen intuicyjności słownika gestów wspiera słuszność pierwszej tezy rozprawy.**

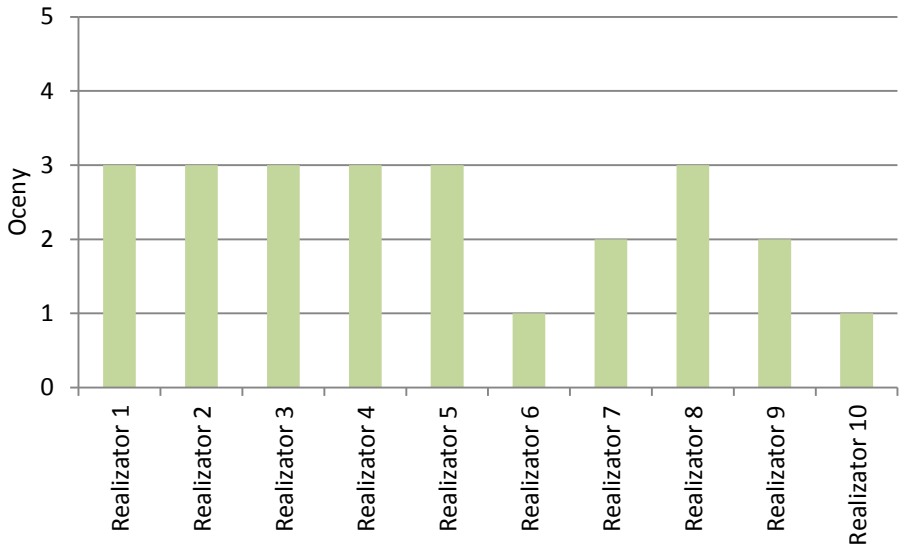


Rys. 6.7 Oceny intuicyjności gestów nadane przez realizatorów

Ocena wygody obsługi

Oceny wygody obsługi, nadane przez poszczególnych realizatorów, przedstawiono na rys. 6.8. Obserwacje dokonane przez autora rozprawy w trakcie pracy realizatorów z systemem pozwalają stwierdzić, że niskie oceny wygody obsługi wynikały w głównej mierze z dwóch czynników. Pierwszym z nich było zmęczenie wynikające z faktu, że interakcja za pomocą gestów, w przeciwieństwie do interakcji z użyciem myszy czy klawiatury, angażuje ręce w sposób uniemożliwiający ich swobodny spoczynek. Drugi czynnik związany był z czasem oczekiwania występującym po uformowaniu określonego kształtu dłoni, poprzedzającym moment prawidłowego wykrycia gestu. Czas ten wynosił około 1,5 s i spowodowany był zastosowaniem bufora uśredniające-

go o długości 50 elementów, w którym gromadzone są wartości prawdopodobieństw przynależności kształtu dłoni do danej klasy gestu. Ograniczenie bufora spowodowałoby skrócenie czasu oczekiwania na rozpoznanie gestu kosztem zaniżenia skuteczności akwizycji. Jeden z realizatorów (realizator nr 8) odnotował w ankiecie uwagę, iż usprawnienie systemu w tym aspekcie uczyniłoby miksowanie za pomocą gestów bardzo wygodnym.

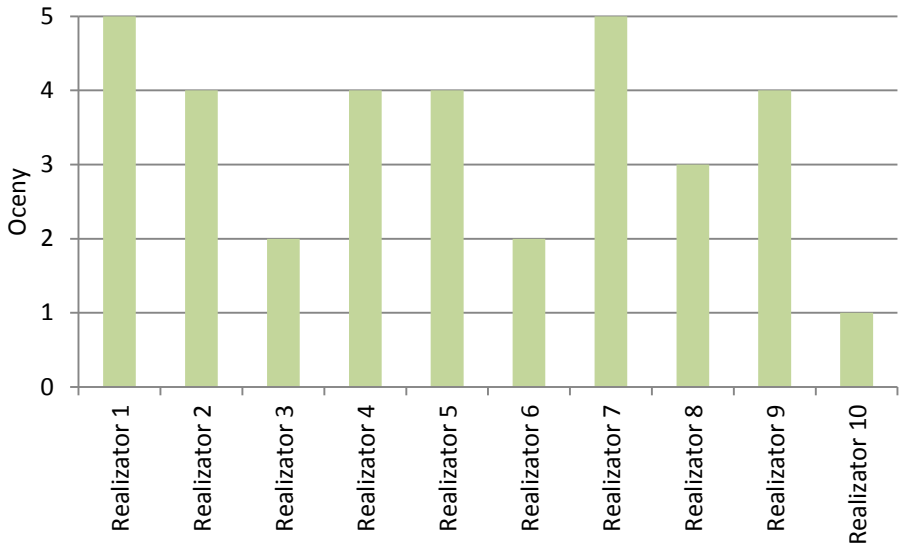


Rys. 6.8 Oceny wygody obsługi systemu za pomocą gestów nadane przez realizatorów

Ocena dokładności edycji parametrów

Oceny dokładności edycji wartości parametrów przedstawiono na rys. 6.9. Na podstawie ankiet wypełnionych przez realizatorów stwierdzić można, że dokładność jaką zapewnia system przedstawiony w rozprawie została oceniona podobnie, jak dla systemu DAW, a w szczególności środowiska Cubase, którym nadzorował opracowany interfejs sterowania za pomocą gestów. Podczas pracy z interfejsem w trybie obsługi za pomocą gestów problematyczne okazało się jednak zatwierdzanie ustalonej wartości parametru. Zmiana kształtu dłoni w celu zakończenia edycji wpływała bowiem na zmianę położenia punktu lokalizacji dłoni w obrazie. W związku z opóźnieniem momentu rozpoznania nowego gestu powodowało to niewielką zmianę oczekiwanej wartości parametru. Aspekt ten stanowił utrudnienie dla ośmiu realizatorów i wpłynął na zaniżenie ocen dokładności. Dwie osoby (realizatorzy nr 1 i 7) potrafiły odpowiednio

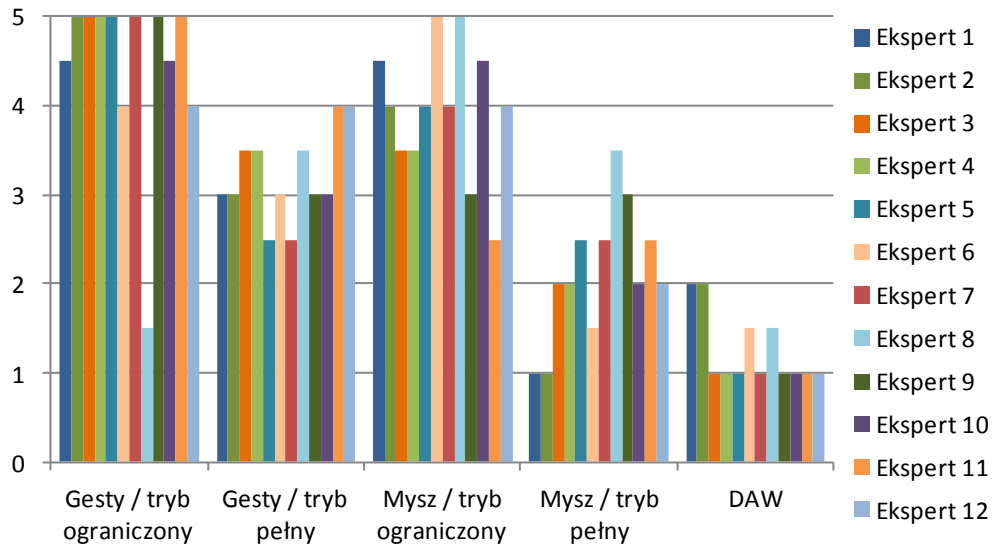
kompensować zmianę położenia dłoni w obrazie w trakcie zmiany gestu poprzez niewielki ruch ręką, przez co, byli oni w stanie za każdym razem ustalić wartość parametru w pełni zgodną z intencjami. Wśród ośmiu realizatorów, którzy nie przyznali maksymalnej oceny, sześciu oceniło przynajmniej jedno z dwóch zgrań uzyskanych w trakcie miksowania za pomocą gestów lepiej niż zgrania otrzymane w oparciu o system DAW.



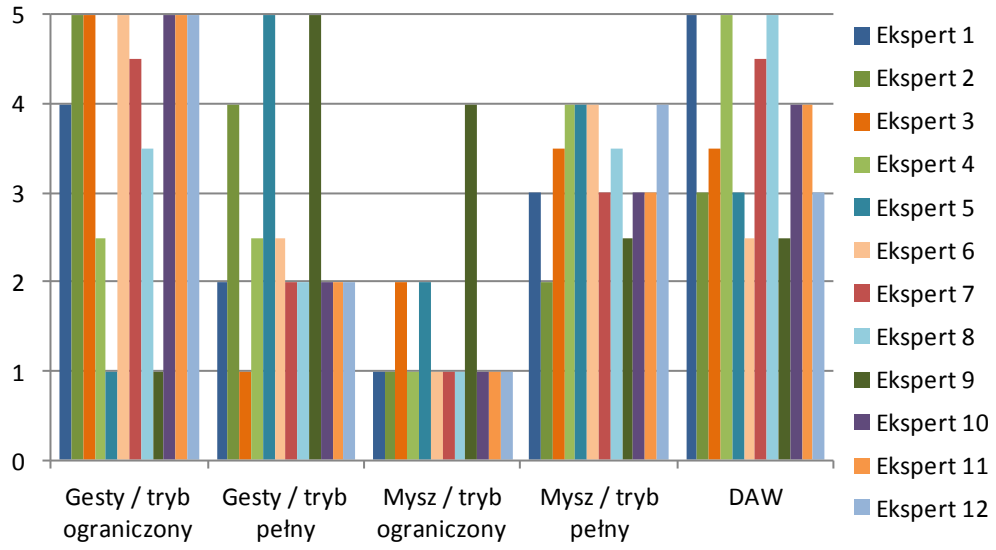
Rys. 6.9 Oceny dokładności edycji wartości parametrów za pomocą gestów nadane przez realizatorów

6.1.5 Analiza wyników testów subiektywnych

Spośród ocen wszystkich słuchaczy, którzy wzięli udział w testach subiektywnych, do analizy wybrano oceny osób, które popełniły maksymalnie 3 błędy, liczba typowań tej samej próbki pomiędzy seriami nie różniła się o więcej niż 1 oraz możliwe było określenie trendu upodobań zgrań na podstawie ocen w obu seriach. Grupa spełniająca powyższe warunki składała się z dwunastu osób. Osoby te w dalszej części rozprawy określane są mianem grupy eksperckiej. W oparciu o analizę typowań w obu seriach dokonano rangowania zgrań zgodnie z pięciostopniową skalą 1 (najgorszy) – 5 (najlepszy). Rozkład rang przedstawiono na rys. 6.10 i 6.11, odpowiednio dla zgrań realizatora nr 3 i 7. Większe dysproporcje w ocenach zgrań realizatora nr 7 wynikały z mniejszych różnic brzmieniowych pomiędzy zgraniami niż w przypadku zgrań realizatora nr 3.

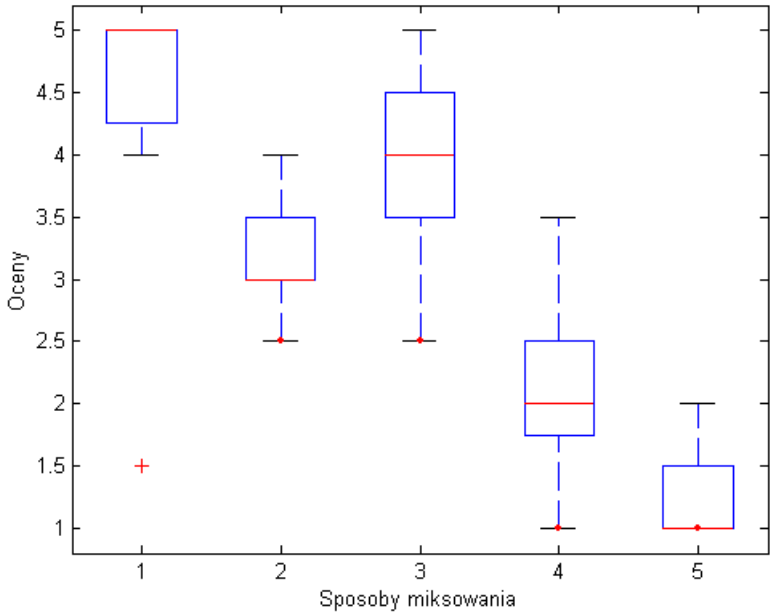


Rys. 6.10 Rozkład rang przydzielonych zgraniom realizatora nr 3 na podstawie ocen subiektywnych z obu serii w teście porównań parami

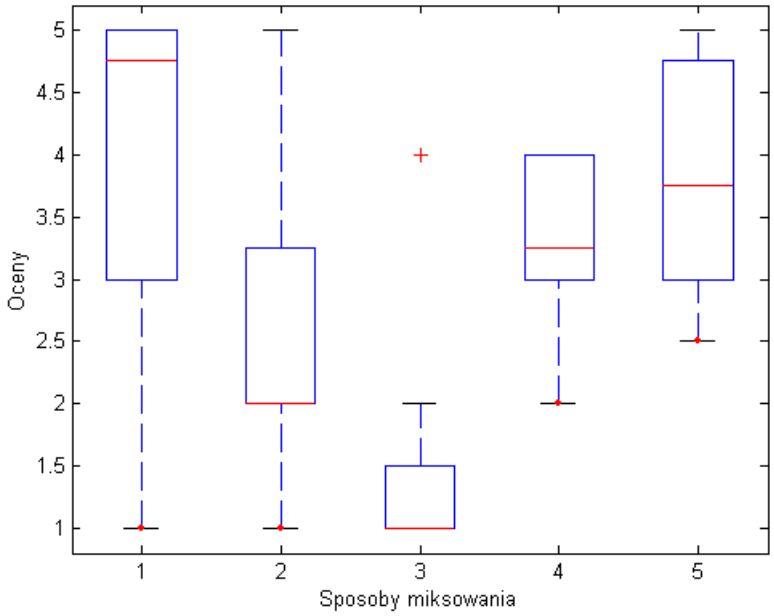


Rys. 6.11 Rozkład rang przydzielonych zgraniom realizatora nr 7 na podstawie ocen subiektywnych z obu serii w teście porównań parami

Analiza rozkładu oceń zgrań została dokonana w oparciu o wykresy pudełkowe (rys. 6.12 i 6.13) oraz testy statystyczne przedstawione w dalszej części niniejszego rozdziału.



Rys. 6.12 Wykres pudełkowy ocen walorów estetycznych zgrań realizatora nr 3 dla różnych sposobów miksowania



Rys. 6.13 Wykres pudełkowy ocen walorów estetycznych zgrań realizatora nr 7 dla różnych sposobów miksowania

Warto zauważyć, że globalny trend ocen ekspertów nie odpowiada ocenom przedzielonym przez realizatorów nr 3 i 7 własnym zgraniom. Szczególnie widoczne jest to

w przypadku zgrań realizatora nr 3 (por. rys. 6.12 i tabela 6.5). Zgranie otrzymane w wyniku miksowania za pomocą gestów w trybie ograniczonego interfejsu graficznego zostało ocenione przez realizatora na „1", podczas gdy w testach subiektywnych uzyskało najwięcej typowań. Powodem takiego stanu rzeczy mogło być zmęczenie słuchu, które pojawiło się u realizatora po wykonaniu miksów. Realizator ten zasugerował w ankiecie ocenę swoich zgrań innym słuchaczom. Trend ocen zgrań przydzielonych przez realizatora nr 7 odbiegał od globalnego trendu ocen ekspertów w ocenie zgrań uzyskanych dla sposobów miksowania ② i ④. Realizator wskazał zgranie otrzymane w wyniku miksowania za pomocą gestów jako lepsze pod względem walorów estetycznych od zgrania uzyskanego w drodze miksowania za pomocą myszy i klawiatury (w obu przypadkach dla trybu pełnego interfejsu graficznego), podczas gdy eksperci ocenili je inaczej.

Otrzymane rozkłady ocen walorów estetycznych zgrań realizatora nr 3 i 7 poddano analizie statystycznej stosując test Friedmana, będący nieparametrycznym odpowiednikiem parametrycznego testu ANOVA (jednoczynnikowej analizy różnic średnich z powtarzalnym pomiarem bądź dwuczynnikowej analizy z klasyfikacją pojedynczą) [10]. Zastosowanie testu parametrycznego nie było możliwe, ponieważ ocena danego zgrania przez eksperta jest zmienną porządkową. Wyniki testów dla ocen przydzielonych zgraniom realizatorów nr 3 i 7 przedstawiono, odpowiednio, w tabelach 6.7 i 6.8.

Tabela 6.7 Zestawienie wartości testu rang Friedmana dla ocen przydzielonych przez ekspertów zgraniom realizatora nr 3 (oznaczenia, jak w tabeli 6.6)

	SS Efekt	df Efekt	MS Efekt	SS Błąd	Df Błąd	MS Błąd	χ^2	*p*
ocena zgrania	83	4	20,75	28	44	0,6364	35,89	0,000

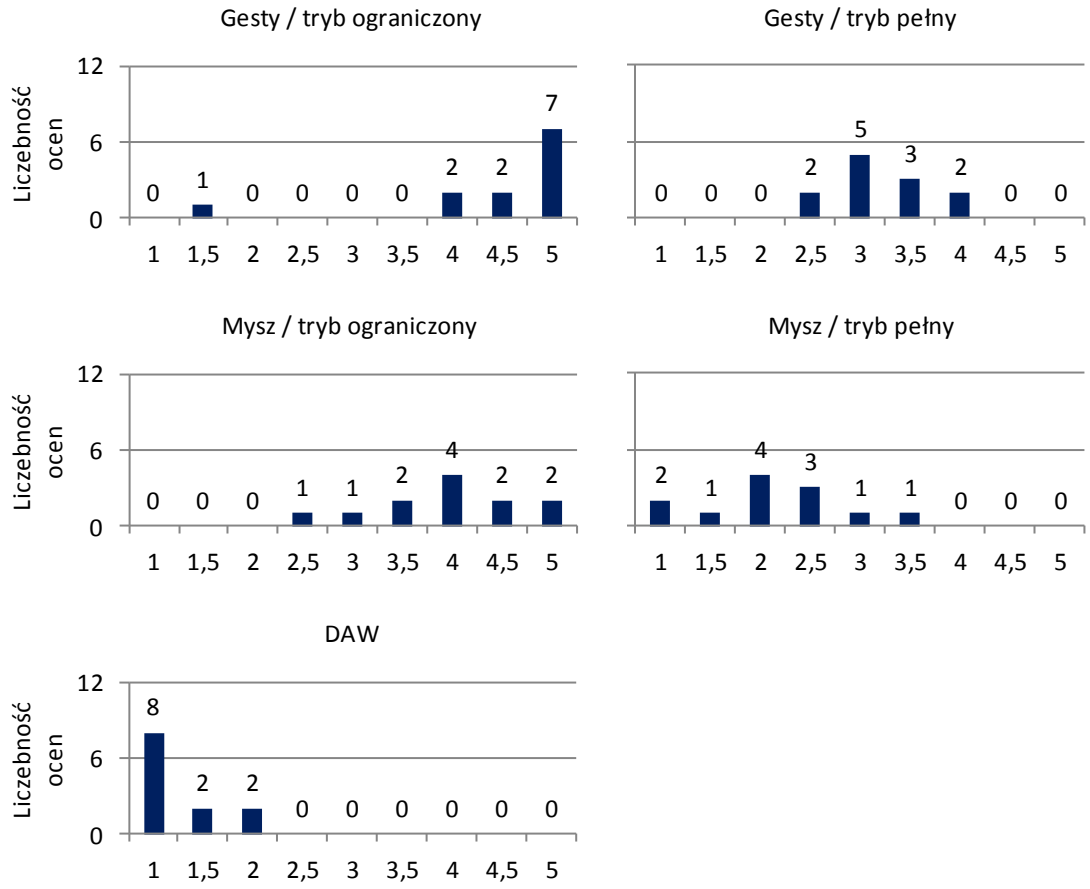
Tabela 6.8 Zestawienie wartości testu rang Friedmana dla ocen przydzielonych przez ekspertów zgraniom realizatora 7(oznaczenia, jak w tabeli 6.6)

	SS Efekt	df Efekt	MS Efekt	SS Błąd	df Błąd	MS Błąd	χ^2	*p*
ocena zgrania	49,4583	4	12,3646	67,5417	44	1,535	20,29	0,0004

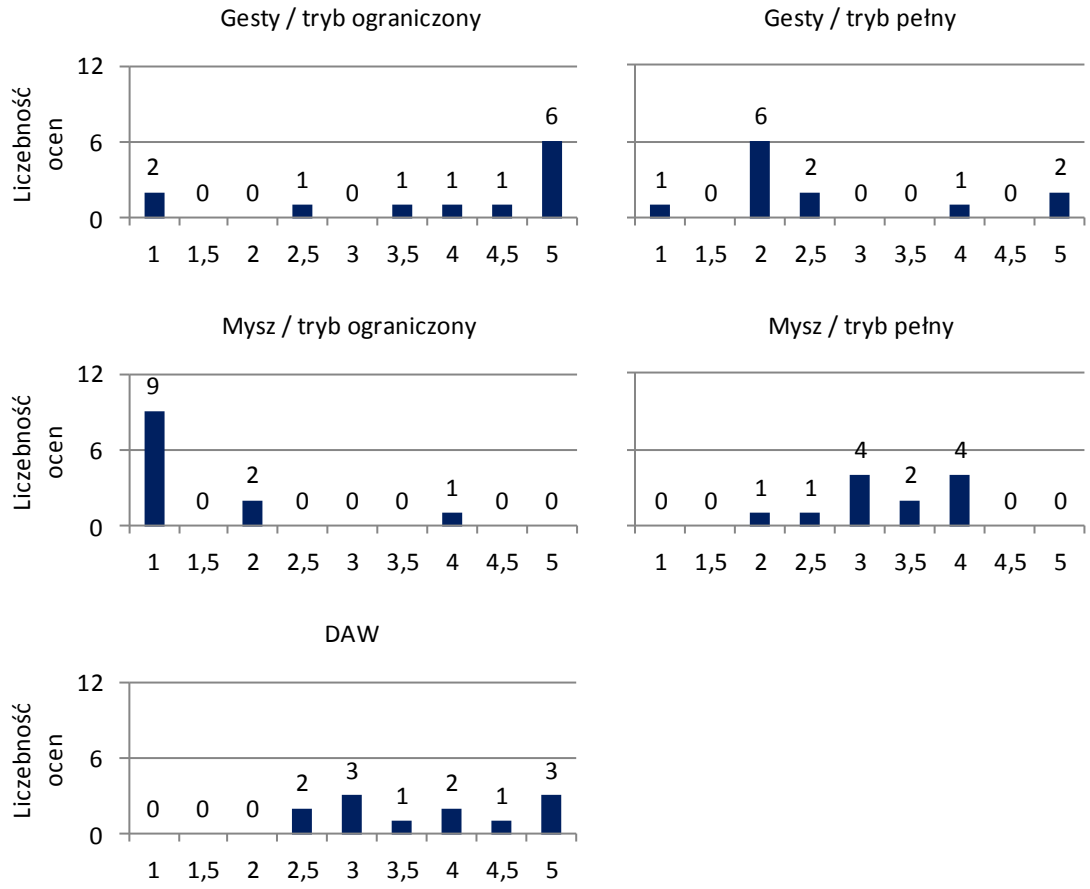
Uzyskane wartości prawdopodobieństw testowych, w obu przypadkach mniejsze niż przyjęty poziom istotności równy 0,05, świadczą o istotnych statystycznie różnicach w rozkładach ocen poszczególnych zgrań. W celu sprawdzenia, pomiędzy którymi parami zgrań wystąpiły istotne statystycznie różnice, wykonano dodatkowo test par rangowanych znaków Wilcoxona. Wyniki testu dla zgrań realizatorów nr 3 i 7 zamieszczono w dodatku E, odpowiednio, w tabelach E.3 i E.4. Wyniki analizy statystycznej rozkładu ocen zgrań realizatora nr 3 odpowiadają wynikom analizy rozkładu wartości parametrów w zakresie wpływu wizualizacji na dobór wartości parametrów. Zarówno dla sposobu interakcji wykorzystującego gesty, jak i mysz komputerową, w zależności od tego czy zastosowano ograniczony interfejs graficzny czy pełny, pomiędzy zgraniami wystąpiły istotne różnice.

Warto zauważyć, że rozkład ocen przydzielonych zgraniom realizatora nr 3 jest bliski rozkładowi określonemu relacją 6.5 w rozdziale 6.1.3. Jedyną różnicą jest niższa ocena zgrań otrzymanych w drodze miksowania bezpośrednio z wykorzystaniem środowiska DAW niż ocena zgrań uzyskanych w procesie miksowania za pomocą myszy w trybie pełnego interfejsu graficznego. Czynnikiem, który to spowodował mógł być odmienny sposób prezentowania informacji w opracowanym systemie niż w środowisku DAW, a w szczególności odzwierciedlanie panoramy i wzmocnienia korektora w lokalizacji kształtu na ekranie. Taki sposób prezentowania i zarządzania informacją wyraźnie był zgodny z upodobaniami realizatora.

Przedstawione na rys. 6.14 i 6.15 histogramy ocen walorów estetycznych zgrań realizatorów 3 i 7, nadanych przez ekspertów, potwierdzają efekt zaobserwowany dla oceny subiektywnej dokonanej przez realizatorów. Mianowicie, żywe brzmienie zgrań otrzymanych wskutek miksowania za pomocą gestów w trybie ograniczonego interfejsu graficznego spowodowało przydzielenie im największej liczby maksymalnych ocen. Jednocześnie, ta sama właściwość była powodem przydzielenia w kilku przypadkach ocen minimalnych. Eksperci, którzy przydzielili oceny minimalne, podobnie jak trzech realizatorów, do których wyników nawiązano w rozdziale 6.1.4, uznali żywe brzmienie za przejaskrawione.



Rys. 6.14 Histogramy ocen walorów estetycznych zgrań realizatora nr 3 dla każdego ze sposobów obsługi procesów miksowania



Rys. 6.15 Histogramy ocen walorów estetycznych zgrań realizatora nr 7 dla każdego ze sposobów obsługi procesów miksowania

6.2 Badanie wydajności systemu

Wydajność opracowanego systemu została zbadana w kontekście czasów wykonania operacji. Badania zostały przeprowadzone na komputerze z procesorem Intel Core 2 Duo P7350 2.0 GHz i pamięcią RAM DDR2 400 MHz o czasach opóźnień 6:6:6:18. Aplikacja pracowała w środowisku 32-bitowego systemu operacyjnego Windows Vista Business. Wybrano ustawienia zaawansowane systemu operacyjnego zapewniające maksymalną wydajność kosztem efektów wizualnych. Ustawiona w systemie rozdzielczość obrazu wynosiła 1024 x 768 pikseli. W każdej iteracji cyklu pracy opracowanego interfejsu z kamery pozyskiwany był obraz o rozmiarze 320 x 240 pikseli. Do takiego samego rozmiaru przeskalowywany był aktualny widok ekranu. W trakcie testów wydajności użytkownik wykonywał gest dynamiczny, który polegał na ruchu ręką do góry i do dołu w zakresie wysokości kadru. Zachowane było stałe tempo wynoszące 1 ruch w danym kierunku na sekundę. Ręka skierowana była zewnętrzną stro-

ną do płaszczyzny ekranu, a palce dłoni były złożone. Kształt ten został zadeklarowany w systemie jako jeden z gestów statycznych. W trakcie przeprowadzania testów bieżący profil zasilania ustawiony był na maksymalną wydajność. W tabeli 6.9 zebrano uśrednione z tysiąca iteracji czasy wykonania kluczowych operacji. Funkcje, dla których w tabeli nie podano nazwy biblioteki, pochodzą z pakietu OpenCV. Opisana w rozdziałach 5.7 i 5.8.1 autorska metoda kalibracji barwowej, nie związana z żadną ze standardowych bibliotek, nie została opatrzona nazwą funkcji w tabeli.

Tabela 6.9 Czasy wykonania poszczególnych operacji

Operacja Nazwa funkcji Czas wykonania [ms]
Pozyskanie ramki cvQueryFrame 9,6335
Operacja Biblioteka Nazwa funkcji Czas wykonania [ms]
Pozyskanie widoku ekranu (w tym*) windef.h, winuser.h, wingdi.h CreateCompatibleBitmap 8,1900
Operacja Nazwa funkcji Czas wykonania [ms]
*Przeskalowanie cvResize 0,2140
Operacja Nazwa funkcji Czas wykonania [ms]
Korekcja perspektywy cvWarpPerspective 6,5546
Operacja Nazwa funkcji Czas wykonania [ms]
Ekstrakcja obszaru efektywnego cvGetSubRect 0,0029
Operacja Nazwa funkcji Czas wykonania [ms]
Odejmowanie obrazów cvAbsDiff 0,2659
Operacja Czas wykonania [ms]
Kalibracja barwowa 3,2800
Operacja Nazwa funkcji Czas wykonania [ms]
Binaryzacja cvThreshold 0,0581
Operacja Nazwa funkcji Czas wykonania [ms]
Filtracja medianowa cvSmooth 3,2844
Operacja Nazwa funkcji Czas wykonania [ms]
Detekcja obiektu w obrazie cvFindContours cvDrawContours 1,7267
Operacja Nazwa funkcji
Filtracja Kalmana cvKalmanCorrect, cvMatMulAdd

Czas wykonania [ms]	0,0050
Operacja Biblioteka Nazwa funkcji Czas wykonania [ms]	Detekcja gestu dynamicznego jFuzzyLogic rule.getDegreeOfSupport 0,0013
Operacja Nazwa funkcji Czas wykonania [ms]	Ekstrakcja kształtu dłoni cvGetSubRect, cvFindContours, cvDrawContours, cvBoundingRect 0,0010
Operacja Biblioteka Nazwa funkcji Czas wykonania [ms]	Detekcja gestu statycznego LIBSVM, OpenCV CalcPGH, svm_predict_probability 0,0003
Całkowity czas wykonania (suma) [ms] Całkowity czas wykonania (zbadany) [ms] Rozdzielczość czasowa [Hz]	**33,22** **43,56** **~22**

Różnica pomiędzy zbadanym całkowitym czasem wykonania pojedynczej iteracji rozpoznawania gestu a czasem wynikającym z sumy wszystkich zmierzonych wartości wynikała z konieczności stosowania dodatkowych operacji, których celem było dostosowanie formatów danych. Przykładem takich operacji mogą być funkcje: *cvFlip*, *cvCvtColor*, *cvSplit*, które realizowały odpowiednio: rotację obrazu względem danej osi, konwersję kolorów, zastąpienie wszystkich składowych koloru jedną, wybraną składową.

Uzyskana rozdzielczość czasowa systemu, wynosząca około 22 klatek na sekundę, pozwoliła na obsługę interfejsu bez zauważalnego opóźnienia. **Ten fakt jest istotny w kontekście udowodnienia pierwszej tezy rozprawy, mówiącej o efektywnym sterowaniu procesami miksowania dźwięku za pomocą gestów.**

6.3 Badanie skuteczności rozpoznawania gestów dynamicznych

W niniejszym rozdziale dokonano analizy porównawczej skuteczności rozpoznawania gestów z wykorzystaniem modułu wnioskowania opartego na ostrych, stałych progach i z wykorzystaniem modułu wnioskowania rozmytego. Przedstawione wyniki stanowią dowód drugiej tezy rozprawy, zgodnie z którą logika rozmyta zastosowana do rozpoznawania gestów dynamicznych, których trajektorią ruchu jest okrąg, pozwala na ich interpretację z wysoką skutecznością.

W testach wzięło udział 20 osób – studentów specjalności Inżynieria Dźwięku i Obrazu w Katedrze Systemów Multimedialnych. Każda osoba proszona była o wyko-

nanie 18 powtórzeń każdego z predefiniowanych gestów dynamicznych. Spośród tych 18 powtórzeń do badań wybrano 10 środkowych gestów. Miało to na celu wyeliminowanie błędów akwizycji, wynikających z nieprawidłowego wykonania gestu w fazie jego nauki lub z powodu zmęczenia. Nie wprowadzono żadnych ograniczeń względem sposobu wykonywania gestów przez osoby, które wzięły udział w testach, takich jak, na przykład, prowadzenie ręki w równej linii lub formowanie określonego kształtu dłoni. Ze względu na przyjętą metodę rozpoznawania gestów dynamicznych, w obrębie pojedynczego ruchu możliwe było wyodrębnienie wielu podstawowych gestów należących do tej samej klasy. Wszystkie te gesty wzięto pod uwagę w analizie skuteczności.

Aby sprawdzić zasadność stosowania logiki rozmytej w rozpoznawaniu gestów dynamicznych, skuteczność rozpoznawania gestów zbadano również dla systemu z modułem detekcji bazującym na ostrych progach. W tabeli 6.10 zawarto wyniki dla gestów obu rąk. W tabelach 6.11 i 6.12 przedstawiono wyniki badania skuteczności dla gestów polegających na ruchu w lewo i w prawo. Brak kolumny odpowiadającej w tabeli danej klasie gestu oznacza, że żaden z gestów zawartych w wierszach nie był błędnie przypisany do tej klasy gestu (wszystkie wartości równe 0,0). Pogrubioną czcionką zaznaczono większe wartości z analizy porównawczej. Ponieważ dane nie miały charakteru rozkładu normalnego ($p < 10^{-9}$ w teście Kołmogorowa-Smirnowa), analizę statystyczną przeprowadzono z wykorzystaniem testu rang Friedmana. Dla przypadków, w których test rang Friedmana pozwolił na przyjęcie hipotezy alternatywnej ($p < 0{,}05$) stanowiącej o istotnym statystycznie zysku z zastosowania logiki rozmytej wykonano dodatkowo test *post hoc* par rangowanych znaków Wilcoxona. Wyniki analizy przedstawiono w sekcji F dodatków do rozprawy.

Tabela 6.10 Skuteczność rozpoznawania gestów obu rąk przy zastosowaniu odpowiednio wnioskowania opartego na ostrych progach (bez logiki rozmytej) i wnioskowania rozmytego (z logiką rozmytą) [%]

G1 – góra, G2 – dół, G3 – powiększenie, G4 – pomniejszenie, G5 – obrót w lewo, G6 – obrót w prawo, G7 – brak gestu

	Bez logiki rozmytej							Z logiką rozmytą					
	G1	G2	G3	G4	G5	G6	G7	G1	G2	G3	G4	G5	G6
G3	0,0	0,0	98,6	0,0	0,0	0,0	1,4	0,2	0,1	**99,7**	0,0	0,0	0,0
G4	0,0	0,2	0,0	98,9	0,0	0,0	0,9	0,6	0,2	0,0	**99,2**	0,0	0,0
G5	0,4	0,0	0,0	0,0	98,2	0,0	1,4	0,5	0,0	0,0	0,4	**99,2**	0,0
G6	0,0	0,2	0,0	0,0	0,0	99,0	0,8	0,2	0,3	0,0	0,4	0,0	**99,1**

Tabela 6.11 Skuteczność rozpoznawania gestów lewej ręki, polegających na ruchu w lewo lub w prawo, przy zastosowaniu odpowiednio wnioskowania opartego na ostrych progach (bez logiki rozmytej) i wnioskowania rozmytego (z logiką rozmytą) [%]

	Bez logiki rozmytej				Z logiką rozmytą				
	lewo	prawo	góra	dół	lewo	prawo	góra	dół	brak
lewo	90,1	0,0	7,9	2,0	**94,5**	0,0	2,4	3,0	0,2
prawo	0,0	90,4	2,8	6,7	0,0	**95,1**	2,4	2,2	0,3

Tabela 6.12 Skuteczność rozpoznawania gestów prawej ręki, polegających na ruchu w lewo lub w prawo, przy zastosowaniu odpowiednio wnioskowania opartego na stałych progach (bez logiki rozmytej) i wnioskowania rozmytego (z logiką rozmytą) [%]

	Bez logiki rozmytej				Z logiką rozmytą			
	lewo	prawo	góra	dół	lewo	prawo	góra	dół
lewo	89,0	0,0	1,9	9,2	**95,6**	0,0	2,2	2,2
prawo	0,0	88,7	8,9	2,4	0,0	**93,4**	3,4	3,2

Zastosowanie zbiorów rozmytych definiujących prędkość ruchu pozwoliło osiągnąć wyższą skuteczność detekcji gestów charakteryzujących się powolnymi ruchami, tj. gestów obu rąk (tabela 6.10). W rozwiązaniu opartym na ostrych progach prędkości i kierunków prędkość powolnego ruchu osiągała wartość mniejszą od ustalonego progu, co uniemożliwiało rozpoznanie gestu. W jednym przypadku zaobserwowano, że zastosowanie logiki rozmytej spowodowało zwiększenie skuteczności rozpoznawania gestów oznaczonych etykietą *bezruch*. Ze względu na drżenie ręki zastosowanie stałych progów uniemożliwiało w tym przypadku prawidłowe rozpoznanie takiego gestu.

Analiza wyników zawartych w tabelach 6.11 i 6.12 oraz w sekcji F dodatków do rozprawy, **pozwala stwierdzić, że zastosowanie systemu wnioskowania opartego na regułach rozmytych przyniosło korzyści w postaci zwiększenia skuteczności rozpoznawania gestów, których trajektoria ruchu nie była wyrównana. Gestami tymi było wykonywane po okręgu przemieszczenie ręki z lewej do prawej strony. Obserwacja ta jest bezpośrednio związana z dowodem drugiej tezy rozprawy.**

Dodatkowo zastosowanie logiki rozmytej poprzez omówione w rozdziale 5.8.3 modelowanie ruchów z uwzględnieniem ich naturalności, zwalnia użytkownika z konieczności zachowywania dbałości o powtarzalność ruchów. Uzyskane wyniki zostały opublikowane w jednej z prac autora niniejszej rozprawy [77].

7 Podsumowanie i wnioski

W rozprawie przedstawiono opracowany przez autora rozprawy system miksowania dźwięku za pomocą gestów rąk. Jedną z motywacji opracowania tego typu systemu było poszerzenie zastosowań systemu rozpoznawania gestów rąk opracowanego wcześniej przez autora rozprawy o interakcję w procesie miksowania sygnałów fonicznych. Dodatkowym bodźcem do przygotowania systemu były doniesienia inżynierów dźwięku, twierdzących, że uzależnienie procesów miksowania od informacji wizualnej w oprogramowaniu DAW jest przyczyną otrzymywania zgrań słabszych pod względem walorów estetycznych od zgrań uzyskiwanych w tradycyjnym podejściu wykorzystującym stół mikserski. Zastosowanie interakcji za pomocą gestów w procesie miksowania dźwięku pozwoliło na stworzenie warunków odpowiednich do sprawdzenia tej hipotezy.

W niniejszej rozprawie nowatorskim aspektem zastosowania gestów rąk w procesie miksowania dźwięku jest możliwość uzyskania zgrania przy ograniczeniu angażowania zmysłu wzroku. Dzięki temu możliwe było opracowanie wiarygodnej metodyki badawczej i zbadanie faktycznego wpływu obrazu na decyzje podejmowane podczas miksowania. W ramach badań sprawdzono cztery konfiguracje opracowanego systemu, różniące się sposobem interakcji (gesty / mysz) i trybem interfejsu graficznego (pełny / ograniczony). Przeprowadzono również analizy porównawcze z oprogramowaniem DAW w postaci aplikacji Steinberg Cubase Studio 5. Zgrania otrzymane w procesie miksowania z wykorzystaniem obu systemów i wszystkich konfiguracji zostały ocenione indywidualnie przez zaproszonych do badań realizatorów nagrań i niezależne grono osób, spośród których wyodrębniono grupę ekspertów. Realizatorzy byli dodatkowo proszeni o ocenę w badaniu ankietowym parametrów takich, jak: intuicyjność słownika gestów, dokładność edycji parametrów, wygoda obsługi. Otrzymany rozkład ocen walorów estetycznych przydzielonych przez realizatorów został poddany analizie statystycznej, do której wykorzystano test rang Friedmana. Uzyskana wartość prawdopodobieństwa testowego (0,7745), dużo większa od przyjętego poziomu istotności, świadczy o braku istotnych statystycznie różnic w walorach estetycznych poszczególnych zgrań. Oznacza to w szczególności, że przeprowadzenie procesów miksowania dźwięku całkowicie za pomocą gestów rąk nie wpłynęło na pogorszenie ich wyników w stosunku

do wyników miksowania z wykorzystaniem środowiska DAW obsługiwanego za pomocą myszy komputerowej, klawiatury i kontrolera MIDI. Dodatkowo przyjęty model rozpoznawania gestów dynamicznych, oparty na logice rozmytej, umożliwił uzyskanie wysokiej wydajności systemu, a mianowicie: 22 klatki/s. Biorąc również pod uwagę zadowalające wyniki badania intuicyjności zastosowanego w systemie słownika gestów, dokładności i wygody obsługi, w konsekwencji można stwierdzić, że udowodniono pierwszą postawioną tezę:

1. **Możliwe jest efektywne sterowanie procesami miksowania dźwięku za pomocą gestów interpretowanych przez komputerowy system analizy obrazu wizyjnego**

W zakresie problemu rozpoznawania gestów, nowatorskim aspektem rozprawy jest połączenie metod przetwarzania obrazu i klasyfikacji obiektów i zdarzeń, realizujące detekcję na tle zmiennego obrazu wyświetlanego przez projektor multimedialny. Większość rozwiązań realizuje detekcję gestów bezpośrednio w strumieniu wizyjnym zawierającym obraz całego użytkownika wykonującego gesty. W celu zapewnienia wysokiej skuteczności detekcji w takich warunkach, wykorzystuje się dodatkowo emitery i czujniki promieniowania podczerwonego bądź kolorowe rękawiczki. W systemie przedstawionym w niniejszej rozprawie wykorzystano obecność projektora multimedialnego tworząc model detekcji gestów, w którym w oparciu o analizę porównawczą odpowiednio przetworzonych strumieni wizyjnych – wyświetlanych przez projektor i pozyskanych z kamery – uzyskano wysoką skuteczność bez konieczności stosowania wyżej wymienionych elementów. W badaniach sprawdzono zasadność stosowania logiki rozmytej do modelowania gestów dynamicznych, porównując skuteczność detekcji gestów w systemie stosującym wnioskowanie rozmyte i w systemie bazującym na ostrych, stałych progach. Dla gestów polegających na ruchu w lewo i w prawo zarówno lewej, jak i prawej ręki, po zastosowaniu logiki rozmytej średnia skuteczność detekcji zwiększyła się. Zgodnie z wynikami przedstawionymi w rozdziale 6.3, dla lewej ręki wyniosła ona 94,5% i 95,1% odpowiednio dla ruchu w lewo i w prawo, co stanowiło zysk odpowiednio 4,4% i 4,7%. Dla ręki prawej uzyskano skuteczność 95,6% i 93,4%,

otrzymując zysk w stosunku do wnioskowania opartego na ostrych progach wynoszący odpowiednio: 6,6% i 4,7%. Wyniki te zostały potwierdzone analizą statystyczną z zastosowaniem testu rang Friedmana i testu par rangowanych znaków Wilcoxona. Dowodzi to drugiej z postawionych tez:

2. Zastosowanie logiki rozmytej w procesie rozpoznawania gestów dynamicznych, dla których trajektorią ruchu jest okrąg, pozwala na ich interpretację z wysoką skutecznością.

Poza oceną walorów estetycznych zgrań, zbadano również wpływ sposobu interakcji (gesty / mysz) oraz wpływ obecności i braku informacji wizualnej odzwierciedlającej wartości parametrów na ich rozkład. Otrzymane wizualizacje miksów uwidoczniły wyraźne różnice w rozmieszczeniu źródeł sygnałów pomiędzy sposobami miksowania, w których wartości parametrów są reprezentowane graficznie i sposobami, w których tryb graficzny był ograniczony. W szczególności zaobserwowano, że miksowanie wsparte pełną informacją wizualną skutkowało większym rozsunięciem źródeł na ekranie. Odpowiadało to szerszej bazie stereofonicznej i intensywniejszemu użyciu korektora częstotliwości. Różnice w rozkładzie wartości parametrów poddano analizie statystycznej z wykorzystaniem testu Friedmana i testu par rangowanych znaków Wilcoxona. Otrzymane wartości prawdopodobieństw testowych, mniejsze od przyjętego poziomu istotności, obiektywnie potwierdziły zaobserwowane przemieszczenia źródeł na wizualizacjach miksów oraz dodatkowo pozwoliły na zidentyfikowanie istotnych różnic w rozkładzie wartości parametru *poziom*.

Niniejsza rozprawa jest próbą wyznaczenia nowego kierunku rozwoju metod miksowania dźwięku i kontrolerów programowych aplikacji DAW. Badania przeprowadzone z udziałem realizatorów nagrań pozwalają na stwierdzenie, że obsługa procesów miksowania sygnałów fonicznych za pomocą gestów rąk jest możliwa i może być uznana za efektywną. Warto przy tym podkreślić, że realizatorzy, biorący udział w testach, po raz pierwszy spotkali się opracowanym systemem, co wskazuje na fakt intuicyjności jego obsługi. Stanowi to kolejny krok na drodze zastępowania zaawansowa-

nych i drogich urządzeń studyjnych rozwiązaniami bardziej uniwersalnymi i tańszymi, co sprzyja popularyzowaniu zagadnień produkcji muzycznej. Wyniki badań rozkładu wartości parametrów w zależności od dostępności informacji wizualnych mogą stanowić uzasadnienie hipotezy stawianej przez realizatorów, że silne uzależnienie interakcji z systemami DAW od informacji wyświetlanej na ekranie monitora jest powodem uzyskiwania słabszych wyników ocenianych w kontekście ich walorów estetycznych niż w oparciu o stoły mikserskie. W tej sytuacji szczególnego znaczenia nabiera interakcja za pomocą gestów rąk. Wykorzystanie interakcji w modelu sterowania może pozwolić na znacząco mniejsze angażowanie zmysłu wzroku w trakcie miksowania.

Dodatkowym istotnym elementem rozprawy jest opracowana metoda rozpoznawania gestów wykonywanych na zmiennym tle, które stanowi obraz wyświetlany przez projektor multimedialny. Jak już zauważono na początku niniejszego rozdziału, podejście, w którym kamera skierowana jest na ekran a nie na użytkownika stanowi przypadek odmienny od typowo stosowanych w rozpoznawaniu gestów. Można zatem powiedzieć, że przyjęta metoda stanowi nowy wkład w obszarze rozpoznawania gestów.

Perspektywy kontynuacji badań

Jak zauważono w rozdziale 5.1.1, stopień kontrastu pomiędzy cieniem rąk a grafiką interfejsu warunkuje skuteczność detekcji rąk w obrazie, co przekłada się na skuteczność rozpoznawania gestów. Z tego względu jako kolor elementów interfejsu graficznego wybrano biel. Z punktu widzenia ergonomii interfejsu pożądana jest jednak ciemna kolorystyka, która przy zastosowaniu projektora w mniejszym stopniu męczy wzrok. Dodatkowo, zgodnie z teorią przedstawioną w pracach Włodarskiego [13] [148], jasne światło w przypadku odbicia od ekranu mogłoby powodować zmiany progów wrażliwości słuchowej. Jako rozwiązanie programowe tego problemu można zaproponować tryb pracy z kamerą skierowaną na użytkownika. Należy jednak zauważyć, że o ile kolorystyka interfejsu przestałaby mieć w tym momencie znaczenie, to istotny zacząłby być kolor tła za użytkownikiem. W kontekście algorytmów zastosowanych w przedstawionym systemie problem ten byłby analogiczny do problemu detekcji rąk na tle obrazu z rzutnika. Co więcej, tło za użytkownikiem może być dowolne również pod względem dynamiki zmian. Zgodnie ze stanem wiedzy [96] [127] detekcja i śledzenie dłoni na takim tle, bez wykorzystania emiterów i odbiorników podczerwieni, jest pro-

blemem złożonym, którego rozwiązania najczęściej nie oferują wystarczającej skuteczności. Konieczne byłoby zatem wprowadzenie ograniczenia zakładającego statyczne tło za użytkownikiem. Pożądane byłoby, aby tło to dodatkowo kontrastowało z kolorem skóry dłoni użytkownika. Jeśliby utrzymać w mocy założenie eliminujące użycie manipulatorów i rękawiczek, to przy zastosowaniu popularnych algorytmów detekcji dłoni opartych na modelu koloru skóry, skuteczność działania systemu mogłaby być jednak niewystarczająca. Czynnikami warunkującymi skuteczność byłyby: kolor ubioru użytkownika oraz kolor twarzy. W konfiguracji, w której kamera skierowana jest na użytkownika, obraz dłoni byłby mniejszy niż w konfiguracji obecnej. Rozmiar cienia dłoni jest bowiem większy niż rozmiar samej dłoni. Dzięki temu, w obecnej wersji systemu możliwe jest operowanie na obrazie o rozmiarze 320 x 240 pikseli. Przetwarzanie większych obrazów spowodowałoby zmniejszenie wydajności systemu.

W ramach kontynuacji badań rozważyć można implementację czterozakresowego parametrycznego korektora częstotliwości. Korektor tego typu jest podstawowym narzędziem pracy inżyniera miksu. Realizacja obsługi siedmiu podstawowych parametrów udostępnianych przez system bazuje obecnie na funkcji *QuickControls*, która umożliwia sterowanie maksymalnie ośmioma parametrami dla każdej ze ścieżek. Udostępnienie wspomnianego korektora za pomocą tej funkcji nie jest zatem możliwe, gdyż wymagałoby użycia 12 parametrów (4 parametry dla dobroci filtrów, 4 parametry dla wzmocnienia i 4 dla częstotliwości). Możliwe jest jednak stworzenie współpracującej z opracowanym systemem wtyczki, interpretującej zmiany kierunków ruchu w stałym przedziale czasu jako ustawienia poszczególnych parametrów.

Przedstawiony w rozprawie system oferuje interakcję dwuwymiarową w zakresie wyświetlania grafiki interfejsu i manipulowania za pomocą gestów. Zastosowana technologia nie umożliwia określania odległości rąk od ekranu ani interpretowania ruchów wykonywanych w osi prostopadłej do płaszczyzny ekranu. W ramach rozwoju oprogramowania rozważyć można zastosowanie technologii obrazu trójwymiarowego oraz technologii rozpoznawania gestów z uwzględnieniem wymiaru głębi. Jak podkreślają Berthaut i in. w pracy, przytoczonej w ramach przeglądu dokonanego w rozdziale 4., środowiska zanurzające użytkownika w przestrzeni trójwymiarowej udostępniają dodatkowy wymiar interakcji z materiałem muzycznym [12]. Wymiar ten może być wykorzystywany nie tylko do wizualizowania akcji przypisanych ruchom użytkownika, ale też

do manipulowania obiektami. W systemie przedstawionym w niniejszej rozprawie dodatkowy wymiar mógłby być wykorzystany do ustawiania pogłosu. Odsunięcie od siebie ręki powodowałoby przesunięcie źródła sygnału w głąb przestrzeni trójwymiarowej, co w naturalny sposób odpowiadałoby zwiększeniu pogłosu. Rozwój wskazanych powyżej elementów mógłby pozwolić na uzyskanie dodatkowych funkcjonalności systemu miksowania za pomocą gestów rąk.

Można się również spodziewać, że wskazane kierunki i perspektywy rozwoju tego typu systemów zaowocują w przyszłości powstaniem systemów komercyjnych wykorzystujących rozpoznawanie gestów rąk w procesie miksowania nagrań.

Bibliografia

[1] N. Aboutabit, D. Beautemps, L. Besacier, Hand and Lip Desynchronization Analysis in French Cued Speech: Automatic Temporal Segmentation of Hand Flow, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1, 633-636, 2006.

[2] Apple, witryna produktu Apple iPad, http://www.apple.com/pl/ipad/, dostęp: 07.2012.

[3] V. Athitsos, S. Sclaroff, Estimating 3D Hand Pose from a Cluttered Image, Proc. IEEE Conferenc, on Computer Vision and Pattern Recognition, Wisconsin, 2003.

[4] F. Avanzini, Interactive Sound, Sound to Sense Sense to Sound – A State of the Art in Sound and Music, D. Rocchesso and P. Polotti (red.), Information Society Technologies, 302-345, 2007.

[5] C.L. Bajaj, V. Pascucci, D.R. Schikore, The contour spectrum, Proc. IEEE Visualization 1997, 167-173, 1997.

[6] D. Balakrishna, P. Sailaja, R. Rao, B. Indurkhya, A novel human robot interaction using the Wiimote, Proc. IEEE International Conference on Robotics and Biomimetics (ROBIO), Tianjin, 645-650, 2010.

[7] W. Balin, J. Loviscach, Gestures to Operate DAW Software, Proc. AES 130th Convention, London, 2011.

[8] T. Baudel, M. Beaudouin-Lofan, Charade: remote control of object using free hand gestures, Communications of the ACM, 36, 7, 28-35, 1993.

[9] B. Bauer, H. Hienz, Relevant features for video-based continuous sign language recognition, Proc. IEEE International Conferenceon Automatic Face and Gesture Recognition, 440-445, 2000.

[10] S. Bech, N. Zacharov, Perceptual Audio Evaluation – Theory, Method and Application, West Sussex, John Wiley & Sons, 2006.

[11] J.R. Beck, M. Garcia, M. Zhong, M. Georgiopoulos, A Backward Adjusting Strategy and Optimization of the C4.5 Parameters to Improve C4.5's

Performance, Proc. XXI Artificial Intelligence Research Symposium FLAIRS, Coconut Grove, FL, 35-40, 2008.

[12] F. Berthaut, M. Desainte-Catherine, M. Hachet, Interacting with the 3D Reactive Widgets for Musical Performance, Journal of New Music Research: New Paradigms for Computer Music, 40, 3, 253-263, 2011.

[13] M. Bogdanowicz, Integracja Percepcyjno-Motoryczna, teoria – diagnoza – terapia, wydanie III, Centrum Metodyczne Pomocy Psychologiczno-Pedagogicznej, 2000.

[14] A. Bosch, A. Zisserman, X. Munoz, Image Classification using Random Forests and Ferns, Proc. IEEE 11th International Conference on Computer Vision ICCV, Rio de Janeiro, 1-8, 2007.

[15] Richard Bowden, David Windridge, Timor Kadir, Andrew Zisserman, Michael Brady, A Linguistic Feature Vector for the Visual Interpretation of Sign Language, Proc. European Conference on Computer Vision, 2004.

[16] D. Bowman, D. Koller, L.F. Hodges, Travel in Immersive Virtual Environments: An Evaluation of Viewpoint Motion Control Techniques, 13, 1, 45-52, 1997.

[17] G. Bradski, A. Kaehler, Learning OpenCV: Computer Vision with the OpenCV Library, Sebastopol, O'Reilly, 2008.

[18] M. Brent, Instance-based Learning: Nearest Neighbour with Generalisation, Hamilton, New Zealand: University of Waikato, 1995, praca magisterska.

[19] T. Burger, A. Benoit, A. Caplier, Extracting Static Hand Gestures in Dynamic Context, Proc. International Conference on Image Processing, 2081-2084, 2006.

[20] W. Buxton, R. Hill, P. Rowley, Issues and Techniques in Touch-Sensitive Tablet Input, Siggraph, 19, San Francisco, 215-224, 1985.

[21] C. Cadoz, Musique, Geste Technologie, Cultures Musicales: Les Nouveaux Gestes de la Musique, H. Genevoix and R. De Vivo (red.), Marseille, Parentheses, 47-92, 1999.

[22] R.R. Campbell, Behind the Gear, Tape Op - The Creative Music Recording

Magazine, 81, 12-13, Feb / Mar 2011.

[23] Zhang Can, Wu Jiankang, Tu Guofang, Object Tracking and QOS Control Using Infrared Sensor and Video Cameras, Proc. 2006 IEEE International Conference on Networking, Sensing and Control ICNSC '06, 974-979, 2006.

[24] J.M. Carroll, Conceptualizing a possible discipline of human–computer interaction, Interacting with Computers, 22, 3-12, 2010.

[25] H. Carr, J. Snoeyink, M. van de Panne, Progressive topological simplification using contour trees and local spatial measures, Proc. 15th Western Computer Graphics Symposium, British Columbia, 2004.

[26] Chih-Chung Chang, Chih-Jen Lin, LIBSVM: a Library for Support Vector Machines, Science, 2, 3, 1-39, 2011.

[27] C. Chang, C. Pengwu, Gesture recognition approach for sign language using curvature scale space and hidden markov model, IEEE International Conference on Multimedia and Expo ICME '04, 2, 1187-1190, 2004.

[28] Kaoa Chang-Yi, Fahn Chin-Shyurng, A Human-Machine Interaction Technique: Hand Gesture Recognition Based on Hidden Markov Models with Trajectory of Hand Motion, Procedia Engineering, 15, 3739-3743, 2011.

[29] Jun Cheng, Can Xie, Wei Bian, Dacheng Tao, Feature fusion for 3D hand gesture recognition by learning a shared hidden space, Pattern Recognition Letters, 33, 476-484, 2012.

[30] Chrome Plugins. http://www.chromeplugins.org/extensions/chrome-gestures-google-chrome-mouse-gestures-extension/, dostęp: 07.2012

[31] C. Chubb, L. Olzak, A. Derrington, Second-order processes in vision: introduction, Journal of the Optical Society of America, 18, 9, 2175-2178, 2001.

[32] J. Congleton, Sound Fascination, Tape Op - The Creative Music Recording Magazine, 81, 2/3, 14-18, 2011.

[33] W.J. Conover, Practical Nonparametric Statistics, wydanie III, John Wiley & Sons, 1999.

[34] Y. Cui, D.L. Swets, J. Weng, Learning-based hand sign recognition using

SHOSLIF-m, Proc. Fifth International Conference on Computer Vision, 631-636, 1995.

[35] J. Davis, M. Shah, Visual gesture recognition, Proc. Vision, Image and Signal Processing, 141, 101-106, 1994.

[36] EiS, Gareth Jones radzi: W domu czy w studiu?, Estrada i Studio, 10, 28, 2009.

[37] C.E. Erdem, S. Ulukaya, A. Karaali, A.T. Erdem, Combining Haar Feature and skin color based classifiers for face detection, Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1497-1500, 2011.

[38] A.F. Everest, Podręcznik akustyki, wydanie IV, W. Kurylak (red.), Katowice, Sonia Draga, 2009.

[39] P. Evjen, J.S. Bradley, S.G. Norcross, The effect of late reflections from above and behind on listener envelopment, Applied Acoustics, 62, 137-153, 2001.

[40] Z. Fangzhou, L. Junshan, Z. Yinghong, Y. Wei, Invariant feature matching based adaptive bandwidth mean shift and its application to infrared object tracking, Proc. 3rd IEEE International Conference on Computer Science and Information technology (ICCSIT), 8, 37-40, 2010.

[41] H. Fillbrandt, S. Akyol, K.-F. Kraiss, Extraction of 3D hand shape and posture from image sequences for sign language recognition, Proc. IEEE International Workshop on Analysis and Modeling of Faces and Gestures, 181-186, 2003.

[42] D. Gibson, The art of mixing: a visual guide to recording engineering production, film, 2003

[43] A. Giegiel, Lubię Wąchać Winyl (LWW), Estrada i Studio, wywiad, 10, 100-104, 2009.

[44] M. Gołaszewska, Zarys Estetyki. Problematyka, metody, teorie, Kraków, W.L., 1973.

[45] M. Gorman, M. Betke, E. Saltzman, A. Lahav, Music Maker - A Camera-

based Music Making Tool for Physical Rehabilitation, Computer Science Technical Report, 2005-032, 2005.

[46] M. Goyani, A. Dhorajiya, R. Paun, Performance Analysis of FDA Based Face Recognition Using Correlation, ANN and SVM, International Journal of Artificial Intelligence and Neural Networks, 1, 1, 108-111, 2011.

[47] Mahesh Goyani, Gunvantsinh Gohil, Amit Chaudhari, Robust Face Recognition in Low Dimensional Subspace Using Reconstructive and Discriminative Features, Proc. International Conference on Communication Systems and Network Technologies CSNT '11, 355-359, 2011.

[48] K.P. Green, Studies of the McGurk effect: implications for theories of speech perception, Proc. Fourth International Conference on Spoken Language (ICSLP), 3, 1652-1655, 1996.

[49] T. Halmrast, Sound Coloration from (Very) Early Reflections, Proc. Acoustical Society of America Chicago 4th june 2001 Conference, ASA, 1-7, 2001.

[50] Y. Hamada, N. Shimada, Y. Shirai, Hand shape estimation using image transition network, humo, 00:161, 2000.

[51] R. Hamilton, q3osc: or how I learned to stop worrying and love the game, Proc. International Computer Music Association Conference, 2008.

[52] T. Heap, F. Samaria, Real-time hand tracking and gesture recognition using smart snakes, proc. Interface to Real and Virtual Worlds, Montpellier, 1-13, 1995.

[53] Cheng Heng-Tze, Chen An Mei, A. Razdan, E. Buller, Contactless gesture recognition system using proximity sensors, Proc. IEEE International Conference on Consumer Electronics (ICCE), 149-150, 2011.

[54] P. Heracleous, Gestures and Lip Shape Integration for Cued Speech Recognition, in 20th International Conference on Pattern Recognition (ICPR), 2238-2241, 2010.

[55] P. Heracleous, D. Beautemps, N. Aboutabit, Cued Speech automatic recognition in normal-hearing and deaf subjects, Speech Communication, 52,

504-512, 2010.

[56] Suk Heung-Il, Sin Bong-Kee, Lee Seong-Whan, Recognizing Hand Gestures using Dynamic Bayesian Network, Proc.8th IEEE International Conference on Automatic Face & Gesture Recognition FG '08, 1-6, 2008.

[57] G. Holmes, A. Donin, I. Witten, WEKA: A Machine Learning Workbench, Proc. Second Australian and New Zealand Conference on Intelligent Information Systems, Brisbane, 357-361, 1994.

[58] Gerard J. Holzmann, Finite State Machines, Design and Validation of Computer Protocols, B.W. Kernighan (red.)., rozdział 8, Prentice-Hall, 162-186, 1991.

[59] Pengyu Hong, Matthew Turk, T.S. Huang, Gesture Modeling and Recognition Using Finite State Machines, Proc. IEEE Conference on Face and Gesture Recognition, 2000.

[60] K. Hoshino, T. Tanimoto, Realtime estimation of human hand posture for robot hand control, Proc. Int. Symposium on Computational Intelligence in Robotics and Automation, Espoo, 99-104, 2005.

[61] C.W. Hsu, C.C. Chang, C.J. Lin, A Practical Guide to Support Vector Classification, Bioinformatics, 1, 1, 1-16, 2010.

[62] C.L. Huang, Wu M.S., A model-based complex background gesture recognition system, Proc. Int. Conference on Systems, Man and Cybernetics, 1, Beijing, 93-98, 1996.

[63] Hunter, Hunter, Estrada i Studio, wywiad, 11, AVT, 88-94, 2009.

[64] A. Jakubik, artykuł na witrynie portalu Psychologia, http://www.psychologia.net.pl/slownik.php?level=30, dostęp 09.2011.

[65] G. Jones. (2007, April) Gareth Jones interview - April 2007.

[66] wywiad z producentem Garethem Jonesem, http://www.youtube.com/watch?v=AXB8dSnNHLc, dostęp 07.2011

[67] Kim Jung-Bae, Park Kwang-Hyun, Bang Won-Chul, Z.Z Bien, Continuous Gesture Recognition System for Korean Sign Language Based on Fuzzy Logic and Hidden Markov Model, Proc. 2002 IEEE International

Conference on Fuzzy Systems (FUZZ-IEEE'02), Honolulu, 1574-1579, 2002.

[68] Kim Jungsoo, He Jiasheng, K Lyons, Thad Starner, The Gesture Watch: A Wireless Contact-free Gesture based Wrist Interface, Proc. 11th IEEE International Symposium on Wearable Computers, Boston, 15 - 22, 2007.

[69] R.E. Kalman, A new approach to linear fi ltering and prediction problems, Journal of Basic Engineering, 82, 35-45, 1960.

[70] Sabine Kastner, A Neural Basis for Human Visual Attention, The Visual Neurosciences, L.M. Chalupa and J.S. Werner (red.), 1, The MIT Press, 2004.

[71] C. Keskin, A. Erkan, L. Akarun, Real Time Hand Tracking and 3D Gesture Recognition for Interactive Interfaces Using HMM, Proc. International Conference on Artificial Neural Networks, 2003.

[72] B. Kostek, Perception-Based Data Processing in Acoustics. Applications to Music Information Retrieval and Psychophysiology of Hearing, Springer Verlag, Series on Cognitive Technologies, Berlin, Heidelberg, New York 2005.

[73] B. Kunka, B. Kostek, Objectivization of audio-video correlation assessment experiments, Proc. 128th Audio Eng. Soc. Convention, preprint no. 8148, London, 2010.

[74] A. Kupryjanow, K. Kaszuba, A Czyżewski, Influence of accelerometer signal pre-processing and classification method on human activity recognition, Elektronika, 3, 2010.

[75] A. Kupryjanow, B. Kunka, B. Kostek, UPDRS tests for Diagnosis of Parkinson's Disease Employing Virtual-Touchpad, Proc. 4th International Workshop on Management and Interaction with Multimodal Information Content - MIMIC '10, Bilbao, 132-136, 2010.

[76] E. Ladavas, N. Bolognini, F. Frassinetti, Multisensory Integration of Audiovisual Inputs in Individuals with and without Visuospatial Impairment, Cognitive Neuroscience of Attention, M.I. Posner (red.), The Guilford Press, 381-392, 2004.

[77] M. Lech, B. Kostek, Fuzzy Rule-based Dynamic Gesture Recognition

Employing Camera & Multimedia Projector, Advances in Intelligent and Soft Computing, Advances in Multimedia and Network Information System Technologies, 80, 69-78, 2010.

[78] M. Lech, B. Kostek, Gesture Controlled Interactive Whiteboard Based on SVM and Fuzzy Logic, Siggraph, Los Angeles, 2010.

[79] M. Lech, B. Kostek, Gesture-based Computer Control System Applied to the Interactive Whiteboard, Proc. 2nd International Conference on Information Technology ICIT'2010, Gdańsk, 75-78, 2010.

[80] M. Lech, B. Kostek, Gesture-based Computer Control System Applied to the Interactive Whiteboard, Zeszyty Naukowe Wydziału ETI PG, 8, 121-126, 2010.

[81] M. Lech, B. Kostek, Hand Gesture Recognition Supported by Fuzzy Rules and Kalman Filters, International Journal of Intelligent Information and Database Systems, 2011.

[82] M. Lech and B. Kostek, The evaluation of influence of ergonomics and multimodal perception on sound mixing results employing the novel gesture-based mixing interface, Journal of the Audio Engineering Society, 2012 (*w druku*).

[83] M Lech, B Kostek, Wydajność środowisk J2SE i C++ / OpenCV w zagadnieniu sterowania komputerem za pomocą gestów, Metody wytwarzania i zastosowania systemów czasu rzeczywistego, L. Trybus and S. Samolej (red.), Warszawa, WKŁ, 187-196, 2010.

[84] M. Lech, B. Kostek, A. Czyżewski, Rozpoznawanie dynamicznych i statycznych gestów rąk w zastosowaniu do sterowania aplikacjami komputerowymi, Zeszyty Naukowe Wydziału ETI Politechniki Gdańskiej - Wytwarzanie Gier Komputerowych, 10, 2011.

[85] M. Lech, B. Kostek, A. Czyżewski, P. Odya, Gesture Recognition Framework for Multimedia Content Viewer Controlling, Proc. IEEE SPA 2009 Signal Processing: Algorithms, Architectures, Arrangements, Poznań, 2009.

[86] M. Lech, B. Kostek, A. Czyżewski, P Odya, Gesture-based Computer Control

System, Elektronika – Konstrukcje, Technologie, Zastosowania, 3, 2010.

[87] G. Levin, Painterly Interfaces for Audiovisual Performance., rozprawa doktorska, 2000.

[88] R.H. Liang, Continuous Gesture Recognition System for Taiwanese Sign Language, rozprawa doktorska, Taiwan, National Taiwan Univ., 1997.

[89] H. Li, M. Greenspan, Model-based segmentation and recognition of dynamic gestures in continuous video streams, Pattern Recognition, 44, 1614-1628, 2011.

[90] S. Litt, Scott Litt, Tape Op - The Creative Music Recording Magazine, 81, 2/3, 20-25, 2011.

[91] S. Łosowski, Sławomir Łosowski, Estrada i Studio, wywiad, 10, 92-97, 2009.

[92] M.T. Marshall, J. Malloch, M.M. Wanderley, Gesture Control of Sound Spatialization for Live Musical Performance, GestureBased HumanComputer Interaction and Simulation, M. Sales Dias (red.), Berlin, Springer, 227-238, 2009.

[93] C. McCormick. Ergates. http://mccormick.cx/projects/ergates/, dostęp: 10.2011

[94] Microsoft, Witryna sieci Web produktu Xbox. http://www.xbox.com/pl-PL/Kinect, dostęp: 10.2011

[95] S. Mitra, Data Mining in Soft Computing Framework: A Survey, Transactions on Neural Networks, 13, 1, 3-14, 2002.

[96] S. Mitra, T. Acharya, Gesture Recognition: A Survey, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 37, 3, 311-324, 2007.

[97] P. Modler, Interactive Control of Musical Structures by Hand Gestures, Proc. Fifth Brazilian Symposium on Computer Music, Belo Horizonte, 143-150, 1998.

[98] P. Modler, Neural Networks for Mapping Hand Gestures to Sound Synthesis Parameters., Trends in Gestural Control of Music, M.M. Wanderley and M. Battier (red.), Paris, 301-314, 2000.

[99] P. Modler, T. Myatt, A Video System for Recognizing Gestures by Artificial Neural Networks for Expressive Musical Control, Lecture Notes in Computer Science, 2915, 541-548, 2004.

[100] P. Modler, T. Myatt, M. Saup, An experimental set of hand gestures for expressive control of musical parameters in realtime, Proc. 2003 conference on New interfaces for musical expression NIME '03, Singapore, 146-150, 2003.

[101] F. Mokhtarian, A.K. Mackworth, A theory of multiscale, curvature-based shape representation for planar curves, IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(8), 789-805, 1992.

[102] M. Muller, Dynamic Time Warping, Information Retrieval for Music and Motion, Springer, 69-84, 2007.

[103] M. Naef, D. Collicot, A vr interface for collaborative 3d audio performance, Proc. 2006 International Conference on New Interfaces for Musical Expression NIME06, Paris, 57-60, 2006.

[104] Nintendo. (2011) Wii, http://www.wii.com, dostęp: 08.2011

[105] J. Olive and S. Pickles, Fijuu, http://www.fijuu.com, dostęp: 08.2011

[106] J. Olive and S. Pickles, q3apd, http://www.selectparks.net/archive/q3apd.htm, dostęp: 08.2011

[107] Oracle Sun Developer Network (SDN). http://java.sun.com/javase/technologies/desktop/, dostęp: 08.2011

[108] S.J. Ovaska, Fusion of Soft Computing and Hard Computing Techniques: a Review of Applications, Proc. International Conference on Systems, Man, and Cybernetics, 1, 370-375, 1999.

[109] B. Owsinski, The Mixing Engineer's Handbook: Second Edition, Boston: Thomson Course Technology PTR, 2006.

[110] J.P. Papa, A.X. Falcao, Suzuki C.T.N., Supervised Pattern Classification based on Optimum-Path Forest, Intl. Journal of Imaging Systems and Technology, 19, 2, 120-131, 2009.

[111] G.T. Park, Z. Bien, C.S. Lee, W. Jang, J.H. Kim, Real-Time Sign Language

Recognition / Generation for Two-Way Communication, World Automation Congress '98, Albuquerque, 1998.

[112] C.B. Park, S.W. Lee, Real-time 3D pointing gesture recognition for mobile robots with cascade HMM and particle filter, Image and Vision Computing, 29, 51-63, 2011.

[113] M. Pec, P. Strumiłło, Estimation of Interaural Time Difference from Measured Head Related Impulse Responses, Proc. Joint Conference NTAV/SPA, 83-87, 2012.

[114] J. Penne, S. Soutschek, L. Fedorowicz, J. Hornegger, Robust real-time 3D time-of-flight based gesture navigation, Proc. 8th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 1-2, 2008.

[115] M. Polański, Hubert Pietrzykowski - PSPaudioware, EiS, 9, 82-91, 2011.

[116] Presonus, Witryna sieci Web firmy Presonus. http://www.presonus.com/products/SoftwareDetail.aspx?SoftwareId=46, dostęp: 08.2012

[117] P. Pumpuang, Anongnart Srivihok, P. Praneetpolgrang, Comparisons of Classifier Algorithms: Bayesian Network, C4.5, Decision Forest and NBTree for Course Registration Planning Model of Undergraduate Students, in IEEE International Conference on Systems, Man and Cybernetics SMC 2008, Singapore, 3647-3651, 2008.

[118] hasło: miksowanie, Encyklopedia PWN, http://encyklopedia.pwn.pl/haslo.php?id=3941373, dostęp: 10.2012

[119] L.R. Rabiner, A tutorial on Hidden Markov Models and selected applications in speech recognition, Proc. IEEE, 77, 2, 257–286, 1989.

[120] G. Reeb, Sur les points singuliers d'une forme de Pfaff completement integrable ou d'une fonction numerique, Comptes Rendus de l'Academie des Sciences, 222, 847-849, 1946.

[121] J.W. Richards, Digital audio mixing, The Radio and Electronic Engineer, 53, 7/8, 257-264, 1983.

[122] A. Saad, An Overview of Hybrid Soft Computing Techniques for Classifier

Design and Feature Selection, Proc. Eighth International Conference on Hybrid Intelligent Systems, 579-583, 2008.

[123] Saitara, Witryna sieci Web firmy Saitara Software, http://saitarasoftware.com/Site/AC-7_Core_Family.html, dostęp: 06.2012

[124] K. Sang-Bum, H. Kyoung-Soo, R. Hae-Chang, M. Sung Hyon, Some Effective Techniques for Naive Bayes, Proc. IEEE Transactions on Knowledge and Data Engineering, 18, 11, 1457-1466, 2006.

[125] R. Seising, Soft concepts for Soft Computing in soft sciences on 20 years of Soft Computing, Proc. IEEE International Conference on Fuzzy Systems FUZZ, 1-8, 2010.

[126] R. Selfridge, J. Reiss, Interactive Mixing Using Wii Controller, Proc. AES 130th Convention, London, 2011.

[127] Masaki Shimizu, Takeharu Yoshizuka, Hiroyuki Miyamoto, A gesture recognition system using stereo vision and arm model fitting, International Congress Series, 1301, 89-92, 2007.

[128] P. Skulimowski, P. Strumiłło, Refinement of depth from stereo camera ego-motion parameters, Electronic Letters, 44, 12, 729-730, 2008

[129] S. Soutschek, J. Penne, J. Hornegger, J. Kornhuber, 3-D Gesture-Based Scene Navigation in Medical Imaging Applications Using Time-Of-Flight Cameras, Proc. 8th IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 1-6, 2008.

[130] T. Starner, A. Pentland, Real-time American sign language recognition from video using Hidden Markov Models, Proc. International Symposium on Computer Vision, 265-270, 1995.

[131] Steinberg, Witryna sieci Web firmy Steinberg, Podręcznik użytkownika kontrolera CC121 ftp://ftp.steinberg.net/Download/Hardware/CC121/CC121_OperationManual_en.pdf, dostęp: 10.2011

[132] B. Stenger, P.R.S. Mendonca, R. Cipolla, Model-based 3d tracking of an articulated hand, 2001.

[133] D.J. Sturman, D. Zeltzer, A Survey of Glove-based Input, Computer Graphics & Applications, 14, 1, 30-39, 1994.

[134] S. Suzuki, K. Abe, Topological structural analysis of digital binary images by border following, Computer Vision, Graphics and Image Processing, 30, 32-46, 1985.

[135] T. Takagi, M. Sugeno, Fuzzy identifcation of systems and its applications to modeling and control, IEEE transactions on systems, man, and cybernetics, 15, 1, 116-132, 1985.

[136] Tan, Vermeulen, Digital audio tape for data storage, IEEE Spectrum, 26, 10, 34-38, Oct. 1989.

[137] Theageman, Mackie HUI MIDI Protocol - The results of a 2-day reverse-engineering-session, SSEI, 2011.

[138] S. Thrun, W. Burgard, D. Fox, Probabilistic Robotics: Intelligent Robotics and Autonomus Agents, MIT Press, 2005.

[139] J. Triesch, Malsburg C., Classification of hand postures against complex backgrounds using elastic graph matching, Image and Vision Computing, 20, 937-943, 2002.

[140] L. Valbom, Adérito Marcos, WAVE: Sound and music in an immersive environment, Computers & Graphics, 29, 6, 871-881, 2005.

[141] M. Van den Bergh, L. Van Gool, Combining RGB and ToF cameras for real-time 3D hand gesture interaction, Workshop on Applications of Computer Vision WACV, 66-72, 2011.

[142] M. Vlaardingen, Hand Models and Systems for Hand Detection, Shape Recognition and Pose Estimation in Video, 2006.

[143] J. Vroomen, B. de Gelder, Perceptual Effects of Cross-modal Stimulation: Ventriloquism and the Freezing Phenomenon, The handbook of multisensory processes, 3, 4, 1-23, 2004.

[144] J. Wachs, H. Stern, Y. Edan, M. Gillam, C. Feied, A Real-Time Hand Gesture Interface for Medical Visualization Applications, Applications of Soft Computing: Advances in Intelligent and Soft Computing, A. Tiwari (red.),

Berlin, Springer, 36, 153-162, 2006.

[145] J. Wachs et al., Real-Time Hand Gesture Interface for Browsing Medical Images, International Journal of Intelligent Computing in Medical Sciences and Image Processing, 2, 1, 15-25, 2008.

[146] S. Waldherr, Roseli Romero, Sebastian Thrun, A Gesture Based Interface for Human-Robot Interaction, Autonomous Robots, 9, 2, 151-173, 2000.

[147] W. Wang, I. Pollak, C.A. Bouman, M.P. Harper, Classification of Images Using Spatial Random Trees, IEEE/SP 13th Workshop on Statistical Signal Processing, Novosibirsk, 449-452, 2005.

[148] Z. Włodarski, Odbiór treści w procesie uczenia się, PWN, Warszawa 1985.

[149] Y. Wu, J. Lin, T.S. Huang, Analyzing and capturing articulated hand motion in image sequences, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(12), 1910-1922, 2005.

[150] Y. Xiuxin, D. Anh, C & Li, A wearable real-time fall detector based on Naive Bayes classifier, Proc. 23rd Canadian Conference on Electrical and Computer Engineering CCECE, 1-4, 2010.

[151] M. Yeasin, S. Chaudhuri, Visual Understanding of dynamic hand gestures, Pattern Recognition, 33, 1805-1817, 2000.

[152] N. Zacharov, J. Huopaniemi, M. Hamalainen, Round robin subjective evaluation of virtual home theatre sound systems at the AES 16th international conference, Proc. Audio Engineering Society 16th International Conference on Spatial Sound Reproduction, 544-556, 1999.

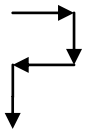
[153] L.A. Zadeh, Fuzzy Sets, Information and Control, 8, 338-353, 1965.

Dodatek A. Wzór ankiety do oceny słownika gestów pod względem intuicyjności

WZÓR ANKIETY

Niniejsza ankieta dotyczy projektu wirtualnej konsolety fonicznej sterowanej za pomocą wyłącznie gestów rąk (statycznych bądź dynamicznych). W poniższej tabeli wymienione i opisane zostały wszystkie gesty zaimplementowane w programie, czyli takie, które program potrafi zinterpretować. Oprócz gestów podstawowych wymienione zostały gesty zgrupowane, czyli konkretna sekwencja gestów podstawowych. Gesty dynamiczne mogą być wykonane dowolnym gestem statycznym (ułożeniem dłoni – 3 pierwsze gesty). Ponadto system rozpoznaje gesty obu rąk, przy czym jedna ręka może wykonywać inny gest niż druga.

	Uformowanie z palców dłoni „gestu OK”
	Skierowanie dłoni wierzchem do kamery
	Uformowanie z palców dłoni litery V
→	Przemieszczenie ręki z lewej do prawej strony kadru
←	Przemieszczenie ręki z prawej do lewej strony kadru
↑	Przemieszczenie ręki z dołu do góry kadru
↓	Przemieszczenie ręki z góry do dołu kadru
← →	Oddalenie wyciągniętych przed siebie rąk, skierowanych wierzchem dłoni do kamery, jedna od drugiej
→ ←	Przybliżenie wyciągniętych przed siebie rąk, skierowanych wierzchem dłoni do kamery, jedna do drugiej
	Wykonanie dynamicznego gestu zgrupowanego (góra-dół)
	Wykonanie dynamicznego gestu zgrupowanego (prawo-lewo)
	Wykonanie dynamicznego gestu zgrupowanego (góra-prawo-dół lub góra-lewo-dół)
	Wykonanie dynamicznego gestu zgrupowanego (dół-prawo-góra lub dół-lewo-góra)
	Wykonanie dynamicznego gestu zgrupowanego (lewo-dół-prawo)

	Wykonanie dynamicznego gestu zgrupowanego (prawo-dół-lewo-dół)
	Wykonanie dynamicznego gestu zgrupowanego (lewo-prawo-dół)
	Wykonanie dynamicznego gestu zgrupowanego (dół-lewo-prawo)
$(\ldots)_2$	Jednoczesne wykonanie sekwencji gestów znajdującej się w nawiasie obiema rękoma

1. Podstawowe informacje: płeć: ☐ mężczyzna ☐ kobieta wiek:
zajęcie: ☐ uczeń ☐ student ☐ inne:

2. W jakim stopniu masz do czynienia z konsoletą foniczną?
☐ zajmuję się tym zawodowo;
☐ zajmuję się tym amatorsko;
☐ zdarza mi się konfigurować konsoletę przy okazji nagrań lub też nagłaśniania jakiegoś wydarzenia;
☐ zapoznałem/am się z konsoletą wyłącznie w ramach studiów;
☐ jedynie widziałem/am, jak ktoś inny konfigurował;
☐ nie miałem/am kontaktu.

3. Zaproponuj gest (bądź sekwencję gestów) z tych dostępnych powyżej dla każdej z podanych funkcji konsolety w poniższej tabeli. Dla ułatwienia pierwsze 3 gesty (statyczne) można zapisać symbolicznie, czyli odpowiednio OK, O, V.

Nazwa funkcji	Opis funkcji	Propozycja gestu
Otwarcie okna wyboru ścieżki	Wyświetlenie okna wyboru ścieżki, na której dokonywane będą dalsze operacje	
Zamknięcie okna wyboru ścieżki	Zamknięcie okna wyboru ścieżki, na której dokonywane będą dalsze operacje	
Odtwarzanie dźwięku	PLAY ⏵	
Zatrzymanie dźwięku	PAUZA ⏸	
Przewijanie w przód	FF ⏩	
Przewijanie w tył	REW ⏪	
Odtwarzanie tylko wybranych	Ustawienie ścieżki w tryb SOLO	
Wyciszenie ścieżki (MUTE)	Ustawienie dla ścieżki 🔇	

Dezaktywacja wszystkich try-	Odtwarzana będzie suma wszystkich niewyciszonych ścieżek	
Dezaktywacja wszystkich try-	Odtwarzana będzie suma wszystkich ścieżek	
Regulacja wywołanego parametru	Zwiększanie bądź zmniejszanie wartości wywołanego wcześniej parametru	
Zmiana poziomu głośności	Wywołanie regulacji poziomu głośności	
Zmiana panoramy	Wywołanie regulacji panoramy	
Dodanie pogłosu (poziom)	Wywołanie regulacji poziomu dźwięku pogłosowego	
Dodanie pogłosu (czas)	Wywołanie regulacji czasu pogłosu	
Stopień kompresji dynamiki	Wywołanie regulacji stopnia kompresji	
Próg kompresji dynamiki	Wywołanie regulacji progu kompresji	
Korekcja częstotliwościowa	Wywołanie wzmacniania/osłabiania pasm częstotliwości	
Przełączenie pomiędzy powyż-	Po wcześniejszym wywołaniu regulacji tych parametrów można przełączać się	

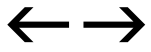
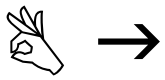
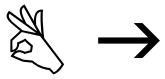
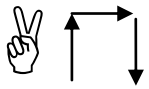
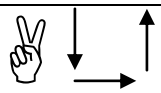
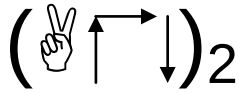
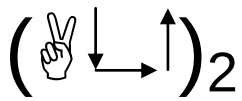
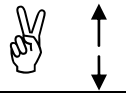
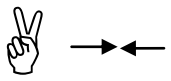
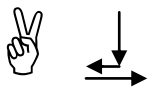
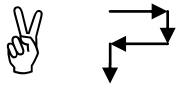
4. Podobnie jak poprzednio zaproponuj gest (bądź sekwencję gestów) dla każdej z podanych funkcji konsolety w poniższej tabeli, jednak tym razem nie ma konieczności posługiwania się słownikiem gestów z pierwszej strony, czyli pełna dowolność.

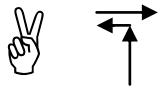
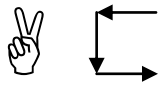
Nazwa funkcji	Opis funkcji	Propozycja gestu
Otwarcie okna wyboru ścieżki	Wyświetlenie okna wyboru ścieżki, na której dokonywane będą dalsze operacje	
Zamknięcie okna wyboru ścieżki	Zamknięcie okna wyboru ścieżki, na której dokonywane będą dalsze operacje	
Odtwarzanie dźwięku	PLAY ⏵	
Zatrzymanie dźwięku	PAUZA ⏸	
Przewijanie w przód	FF ⏩	
Przewijanie w tył	REW ⏪	
Odtwarzanie tylko wybranych	Ustawienie ścieżki w tryb SOLO	
Wyciszenie ścieżki (MUTE)	Ustawienie dla ścieżki 🔇	
Dezaktywacja wszystkich try-	Odtwarzana będzie suma wszystkich niewyciszonych ścieżek	
Dezaktywacja wszystkich try-	Odtwarzana będzie suma wszystkich ścieżek	
Regulacja wywołanego parametru	Zwiększanie bądź zmniejszanie wartości wywołanego wcześniej parametru	
Zmiana poziomu głośności	Wywołanie regulacji poziomu głośności	
Zmiana panoramy	Wywołanie regulacji panoramy	
Dodanie pogłosu (poziom)	Wywołanie regulacji poziomu dźwięku pogłosowego	
Dodanie pogłosu (czas)	Wywołanie regulacji czasu pogłosu	
Stopień kompresji dynamiki	Wywołanie regulacji stopnia kompresji	
Próg kompresji dynamiki	Wywołanie regulacji progu kompresji	
Korekcja częstotliwościowa	Wywołanie wzmacniania/osłabiania pasm częstotliwości	
Przełączenie pomiędzy powyż-	Po wcześniejszym wywołaniu regulacji tych parametrów można przełączać się	

5. Uwagi odnoście funkcjonalności

.....
.....
.....
.....

6. Określ stopień intuicyjności poniższych gestów w kontekście wywoływanych przez nie funkcji w skali od 1 (najsłabsza intuicyjność) do 5 (najlepsza).

Gest	Nazwa funkcji	Opis funkcji	Ocena
	Wybranie ścieżki (I)	Wyświetlenie okna wyboru ścieżki, na której dokonywane będą dalsze operacje	
	Wybranie ścieżki (II)	Zamknięcie okna wyboru ścieżki, na której dokonywane będą dalsze operacje	
	Odtwarzanie dźwięku	Odtwarzanie, jeśli dźwięk akurat nie jest odtwarzany	
	Zatrzymanie dźwięku	Zatrzymanie przy aktywnym odtwarzaniu dźwięku	
	Przewijanie w przód	Przewijanie w przód przy aktywnym odtwarzaniu dźwięku	
	Przewijanie w tył	Przewijanie w tył, jeśli dźwięk nie jest odtwarzany	
	Odtwarzanie tylko wybranych ścieżek (SOLO)	Ustawienie ścieżki w tryb SOLO	
	Wyciszenie ścieżki (MUTE)	Ustawienie dla ścieżki trybu MUTE	
	Dezaktywacja wszystkich trybów SOLO	Odtwarzana będzie suma wszystkich niewyciszonych ścieżek	
	Dezaktywacja wszystkich trybów wyciszenia	Odtwarzana będzie suma wszystkich ścieżek	
	Regulacja wywołanego parametru	Zwiększanie bądź zmniejszanie wartości wywołanego wcześniej parametru	
	Zmiana poziomu albo panoramy	Wywołanie regulacji poziomu głośności albo panoramy	
	Przełączenie pomiędzy ustawianiem poziomu lub panoramy	Po wcześniejszym wywołaniu regulacji tych parametrów powoduje przełączanie się pomiędzy nimi	
	Dodanie pogłosu	Wywołanie regulacji poziomu dźwięku pogłosowego	
	Kompresja dynamiki	Wywołanie regulacji stopnia kompresji	

	Kompresja dynamiki	Wywołanie regulacji progu kompresji	
	Korekcja częstotliwościowa	Wywołanie regulacji wzmacniania pasm częstotliwości	
~	Narysowanie krzywej korekcji	Alternatywny sposób korekcji częstotliwości po wywołaniu w powyższy sposób	

Uwagi:

.....
.....
.....
.....

Dodatek B. Definicja systemu rozmytego w kodzie FCL

```
// Block definition (there may be more than one block per file)
FUNCTION_BLOCK fuzzyGestureSystem

// Define input variables
VAR_INPUT
      velocityLt0 : REAL;
      velocityLt1 : REAL;
      velocityRt0 : REAL;
      velocityRt1 : REAL;
      directionLt0 : REAL;
      directionLt1 : REAL;
      directionRt0 : REAL;
      directionRt1 : REAL;
END_VAR

// Define output variable
VAR_OUTPUT
      gesture : REAL;
END_VAR

FUZZIFY velocityLt0
      TERM vsmall := (0, 1)  (2, 0);
      TERM small  := (0, 1)  (12, 0);
    TERM medium := (8, 0) (30,1) (50,0);
    TERM high   := (40, 0) (100, 1);
END_FUZZIFY

FUZZIFY velocityLt1
      TERM vsmall := (0, 1)  (2, 0);
      TERM small  := (0, 1)  (12, 0);
    TERM medium := (8, 0) (30,1) (50,0);
    TERM high   := (40, 0) (100, 1);
END_FUZZIFY

FUZZIFY velocityRt0
      TERM vsmall := (0, 1)  (2, 0);
      TERM small  := (0, 1)  (12, 0);
    TERM medium := (8, 0) (30,1) (50,0);
    TERM high   := (40, 0) (100, 1);
END_FUZZIFY

FUZZIFY velocityRt1
      TERM vsmall := (0, 1)  (2, 0);
      TERM small  := (0, 1)  (12, 0);
    TERM medium := (8, 0) (30,1) (50,0);
    TERM high   := (40, 0) (100, 1);
END_FUZZIFY

FUZZIFY directionLt0
    TERM North := (0, 1)  (90, 0)  (270, 0) (360, 1);
      TERM East  := (0, 0)   (90, 1)  (180, 0); // gauss 90 30;
      TERM South := (90, 0)  (180, 1) (270, 0); // gauss 180 30;
      TERM West  := (180, 0) (270, 1) (360, 0); // gauss 270 30;
END_FUZZIFY

FUZZIFY directionLt1
    TERM North := (0, 1)  (90, 0)  (270, 0) (360, 1);
      TERM East  := (0, 0)   (90, 1)  (180, 0); // gauss 90 30;
```

```
      TERM South := (90, 0)  (180, 1) (270, 0); // gauss 180 30;
      TERM West  := (180, 0) (270, 1) (360, 0); // gauss 270 30;
END_FUZZIFY

FUZZIFY directionRt0
      TERM North := (0, 1)  (90, 0)  (270, 0) (360, 1);
      TERM East  := (0, 0)   (90, 1)  (180, 0); // gauss 90 30;
      TERM South := (90, 0)  (180, 1) (270, 0); // gauss 180 30;
      TERM West  := (180, 0) (270, 1) (360, 0); // gauss 270 30;
END_FUZZIFY

FUZZIFY directionRt1
      TERM North := (0, 1)  (90, 0)  (270, 0) (360, 1);
      TERM East  := (0, 0)   (90, 1)  (180, 0); // gauss 90 30;
      TERM South := (90, 0)  (180, 1) (270, 0); // gauss 180 30;
      TERM West  := (180, 0) (270, 1) (360, 0); // gauss 270 30;
END_FUZZIFY


// Defuzzify output variable 'gesture'
DEFUZZIFY gesture
      TERM g0 := 0;     // hand(s) steady
      TERM g1 := 1;     // hand left
      TERM g2 := 2;     // hand right
      TERM g3 := 3;     // hand up
      TERM g4 := 4;     // hand down
      TERM g5 := 5;     // zoom out
      TERM g6 := 6;     // zoom in
      TERM g7 := 7;     // rotate left
      TERM g8 := 8;     // rotate right

    METHOD : COGS;
    DEFAULT := 0;
END_DEFUZZIFY

RULEBLOCK No1
    AND : MIN;
    ACCU : MAX;

// both hands steady
RULE 0 : IF velocityLt0 IS vsmall AND velocityLt1 IS vsmall AND
            velocityRt0 IS vsmall AND velocityRt1 IS vsmall
                        THEN gesture IS g0;

// beginning phase of hand movement in the left direction (can be arc
motion) for left hand
RULE 1 : IF directionLt0 IS North AND directionLt1 IS West AND
            velocityLt0 IS NOT small AND velocityLt1 IS NOT small AND
            velocityRt0 IS vsmall AND velocityRt1 IS vsmall
                        THEN gesture IS g1;

// moving hand left (middle phase) for left hand
RULE 2 : IF directionLt0 IS West AND directionLt1 IS West AND
            velocityLt0 IS NOT small AND velocityLt1 IS NOT small AND
            velocityRt0 IS vsmall AND velocityRt1 IS vsmall
                        THEN gesture IS g1;

// ending phase of hand movement in the left direction (can be arc
motion) for left hand
RULE 3 : IF directionLt0 IS West AND directionLt1 IS South AND
            velocityLt0 IS NOT small AND velocityLt1 IS NOT small AND
```

```
            velocityRt0 IS vsmall AND velocityRt1 IS vsmall
                        THEN gesture IS g1;

// beginning phase of hand movement in the left direction (can be arc
motion) for right hand
RULE 4 : IF directionRt0 IS North AND directionRt1 IS West AND
            velocityRt0 IS NOT small AND velocityRt1 IS NOT small AND
            velocityLt0 IS vsmall AND velocityLt1 IS vsmall
                        THEN gesture IS g1;

// moving hand left (middle phase) for right hand
RULE 5 : IF directionRt0 IS West AND directionRt1 IS West AND
            velocityRt0 IS NOT small AND velocityRt1 IS NOT small AND
            velocityLt0 IS vsmall AND velocityLt1 IS vsmall
                        THEN gesture IS g1;

// ending phase of hand movement in the left direction (can be arc
motion) for right hand
RULE 6 : IF directionRt0 IS West AND directionRt1 IS South AND
            velocityRt0 IS NOT small AND velocityRt1 IS NOT small AND
            velocityLt0 IS vsmall AND velocityLt1 IS vsmall
                        THEN gesture IS g1;

// both hands left
// RULE 7 : IF velocityLt1 IS NOT small AND velocityRt1 IS NOT small AND
//             directionLt1 IS West AND directionRt1 IS West
//                         THEN gesture IS g1;

// beginning phase of hand movement in right direction (can be arc
motion) for left hand
RULE 8 : IF directionLt0 IS North AND directionLt1 IS East AND
            velocityLt0 IS NOT small AND velocityLt1 IS NOT small AND
            velocityRt0 IS vsmall AND velocityRt1 IS vsmall
                        THEN gesture IS g2;

// middle phase of hand movement in right direction for left hand
RULE 9 : IF directionLt0 IS East AND directionLt1 IS East AND
            velocityLt0 IS NOT small AND velocityLt1 IS NOT small AND
            velocityRt0 IS vsmall AND velocityRt1 IS vsmall
                        THEN gesture IS g2;

// ending phase of hand movement in right direction for left hand
RULE 10 : IF directionLt0 IS East AND directionLt1 IS South AND
             velocityLt0 IS NOT small AND velocityLt1 IS NOT small AND
             velocityRt0 IS vsmall AND velocityRt1 IS vsmall
                        THEN gesture IS g2;

// beginning phase of hand movement in right direction (can be arc
motion) for right hand
RULE 11 : IF directionRt0 IS North AND directionRt1 IS East AND
             velocityRt0 IS NOT small AND velocityRt1 IS NOT small AND
             velocityLt0 IS vsmall AND velocityLt1 IS vsmall
                        THEN gesture IS g2;

// middle phase of hand movement in right direction for right hand
RULE 12 : IF directionRt0 IS East AND directionRt1 IS East AND
             velocityRt0 IS NOT small AND velocityRt1 IS NOT small AND
             velocityLt0 IS vsmall AND velocityLt1 IS vsmall
                        THEN gesture IS g2;

// ending phase of hand movement in right direction for right hand
```

```
RULE 13 : IF directionRt0 IS East AND directionRt1 IS South AND
             velocityRt0 IS NOT small AND velocityRt1 IS NOT small AND
             velocityLt0 IS vsmall AND velocityLt1 IS vsmall
                       THEN gesture IS g2;

// both hands right
//RULE 14 : IF velocityLt1 IS NOT small AND velocityRt1 IS NOT small AND
//            directionLt1 IS East AND directionRt1 IS East
//                       THEN gesture IS g2;

// left hand up (middle phase)
RULE 16 : IF directionLt0 IS North AND directionLt1 IS North AND
            (velocityLt0 IS NOT small OR velocityLt1 IS NOT small) AND
             velocityRt0 IS vsmall AND velocityRt1 IS vsmall
                       THEN gesture IS g3;

// right hand up (middle phase)
RULE 18 : IF directionRt0 IS North AND directionRt1 IS North AND
            (velocityRt0 IS NOT small OR velocityRt1 IS NOT small) AND
             velocityLt0 IS vsmall AND velocityLt1 IS vsmall
                       THEN gesture IS g3;

// both hands up
// RULE 19 : IF directionLt1 IS North AND directionRt1 IS North AND
//              velocityLt1 IS NOT small AND velocityRt1 IS NOT small
//                       THEN gesture IS g3;

// left hand down
RULE 21 : IF directionLt0 IS South AND directionLt1 IS South AND
            (velocityLt0 IS NOT small OR velocityLt1 IS NOT small) AND
             velocityRt0 IS vsmall AND velocityRt1 IS vsmall
                       THEN gesture IS g4;

// right hand down
RULE 23 : IF directionRt0 IS South AND directionRt1 IS South AND
            (velocityRt0 IS NOT small OR velocityRt1 IS NOT small) AND
             velocityLt0 IS vsmall AND velocityLt1 IS vsmall
                       THEN gesture IS g4;

// both hands down
//RULE 24 : IF (directionLt1 IS South AND directionRt1 IS South) AND
               velocityLt1 IS NOT small AND velocityRt1 IS NOT small
                       THEN gesture IS g4;

// zoom out
RULE 25 : IF directionLt0 IS East AND directionLt1 IS East AND
             directionRt0 IS West AND directionRt1 IS West AND
             velocityLt0 IS NOT vsmall AND velocityLt1 IS NOT vsmall AND
             velocityRt0 IS NOT vsmall AND velocityRt1 IS NOT vsmall
                       THEN gesture IS g5;

// zoom in
RULE 26 : IF directionLt0 IS West AND directionLt1 IS West AND
             directionRt0 IS East AND directionRt1 IS East AND
             velocityLt0 IS NOT vsmall AND velocityLt1 IS NOT vsmall AND
             velocityRt0 IS NOT vsmall AND velocityRt1 IS NOT vsmall
                       THEN gesture IS g6;

// zoom in (left hand starting before border)
RULE 27 : IF directionLt1 IS West AND directionRt0 IS East AND
             directionRt1 IS East AND
```

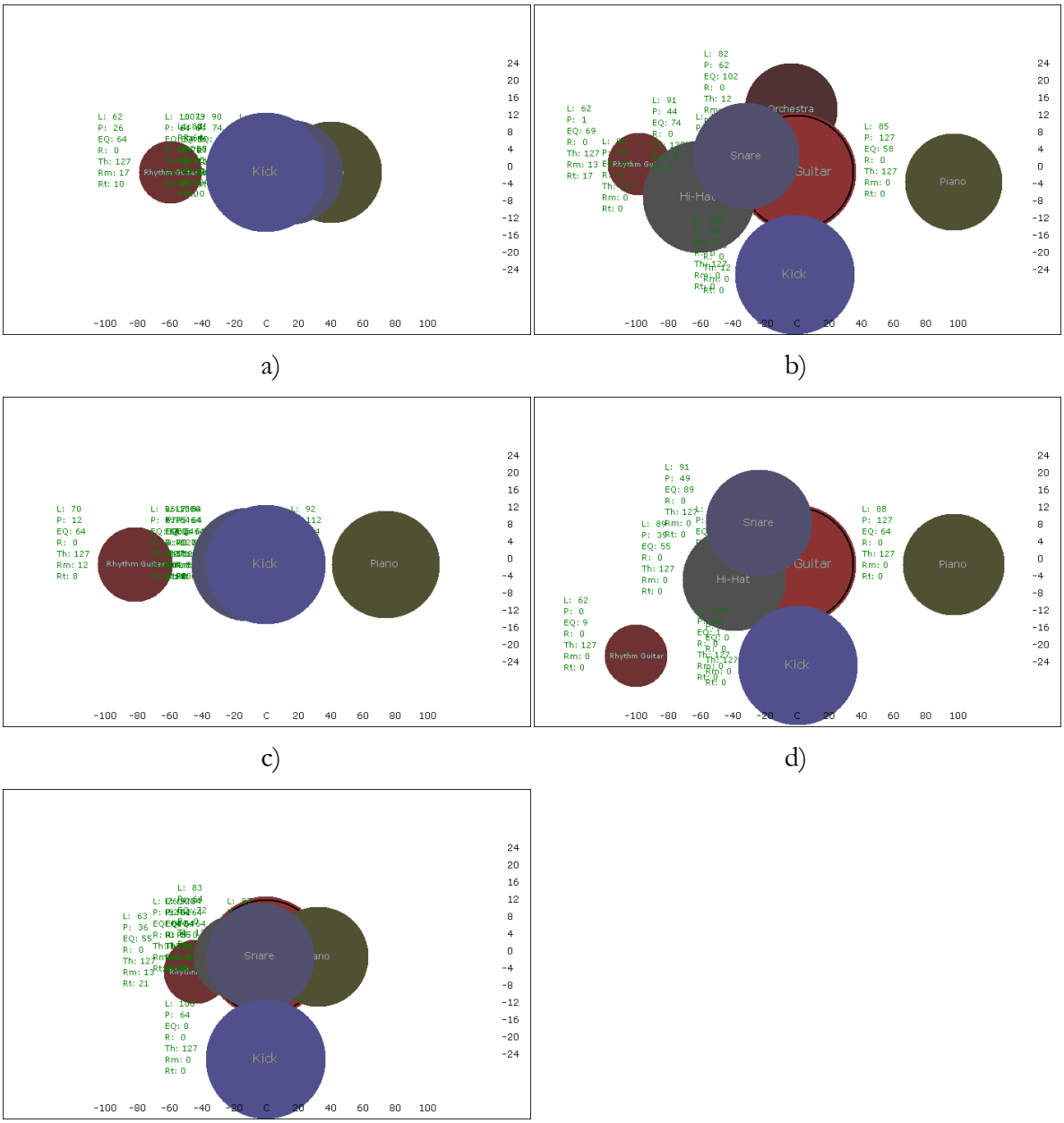
```
            velocityLt1 IS NOT vsmall AND velocityRt0 IS NOT vsmall AND
            velocityRt1 IS NOT vsmall
                        THEN gesture IS g6;

// zoom in (right hand starting before border)
RULE 28 : IF directionLt0 IS West AND directionLt1 IS West AND
            directionRt1 IS East AND
            velocityLt0 IS NOT vsmall AND velocityLt1 IS NOT vsmall AND
            velocityRt1 IS NOT vsmall
                        THEN gesture IS g6;

// rotate left
RULE 29 : IF directionLt0 IS South AND directionLt1 IS South AND
            directionRt0 IS North AND directionRt1 IS North AND
           (velocityLt1 IS NOT vsmall AND velocityLt0 IS NOT vsmall) AND
           (velocityRt1 IS NOT vsmall AND velocityRt0 IS NOT vsmall)
                        THEN gesture IS g7;

// rotate right
RULE 30 : IF directionLt0 IS North AND directionLt1 IS North AND
            directionRt0 IS South AND directionRt1 IS South AND
           (velocityLt1 IS NOT vsmall AND velocityLt0 IS NOT vsmall) AND
           (velocityRt1 IS NOT vsmall AND velocityRt0 IS NOT vsmall)
                        THEN gesture IS g8;

END_RULEBLOCK

END_FUNCTION_BLOCK
```

Dodatek C. Wizualizacje zgrań i wartości parametrów

Ponieważ inżynierowie dźwięku są przyzwyczajeni do pracy ze ścieżkami opisanymi w języku angielskim, w opracowanym systemie wszelkie nazwy i oznaczenia funkcji wyświetlane były w tym języku. Poniżej przedstawiono słownik oznaczeń widocznych na wizualizacjach zgrań oraz w tabelach.

Słownik oznaczeń

nazwa ścieżki (skrót w tabeli) lub oznaczenie parametru	polska nazwa ścieżki lub parametru
Kick	Bęben basowy
Snare	Werbel
Hi-Hat (H-H)	Talerze (Hi-Hat)
Bass	Gitara basowa
Piano	Fortepian
Lead guitar (Lead)	Gitara prowadząca
Rhythm guitar (Rth.)	Gitara rytmiczna
Orchestra (Orch.)	Orkiestra symfoniczna
Lev	Poziom
Pan	Panorama
EQ	Wzmocnienie korektora
Thr	Próg kompresji dynamiki
Rat	Stopień kompresji dynamiki
Mix	Miks pogłosu
Tim	Czas pogłosu



Rys. C.1 Wizualizacje zgrań realizatora 1: (a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

Tabela C.1 Wartości kontrolerów parametrów zgrań realizatora 1

(a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

a)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	64	127	0	0	0
Snare	90	74	64	127	0	0	0
H-H	75	64	58	127	0	0	0
Bass	83	64	58	127	0	0	0
Piano	88	90	64	127	0	0	0
Lead	100	64	64	85	127	5	45
Rth.	62	26	64	127	0	17	10
Orch.	79	64	64	127	0	0	0

b)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	63	0	127	0	0	0
Snare	91	44	74	127	0	0	0
H-H	95	25	49	127	0	0	0
Bass	83	62	0	127	0	0	0
Piano	85	127	58	127	0	0	0
Lead	100	64	64	103	51	13	78
Rth.	62	1	69	127	0	13	17
Orch.	82	62	102	127	0	0	0

c)

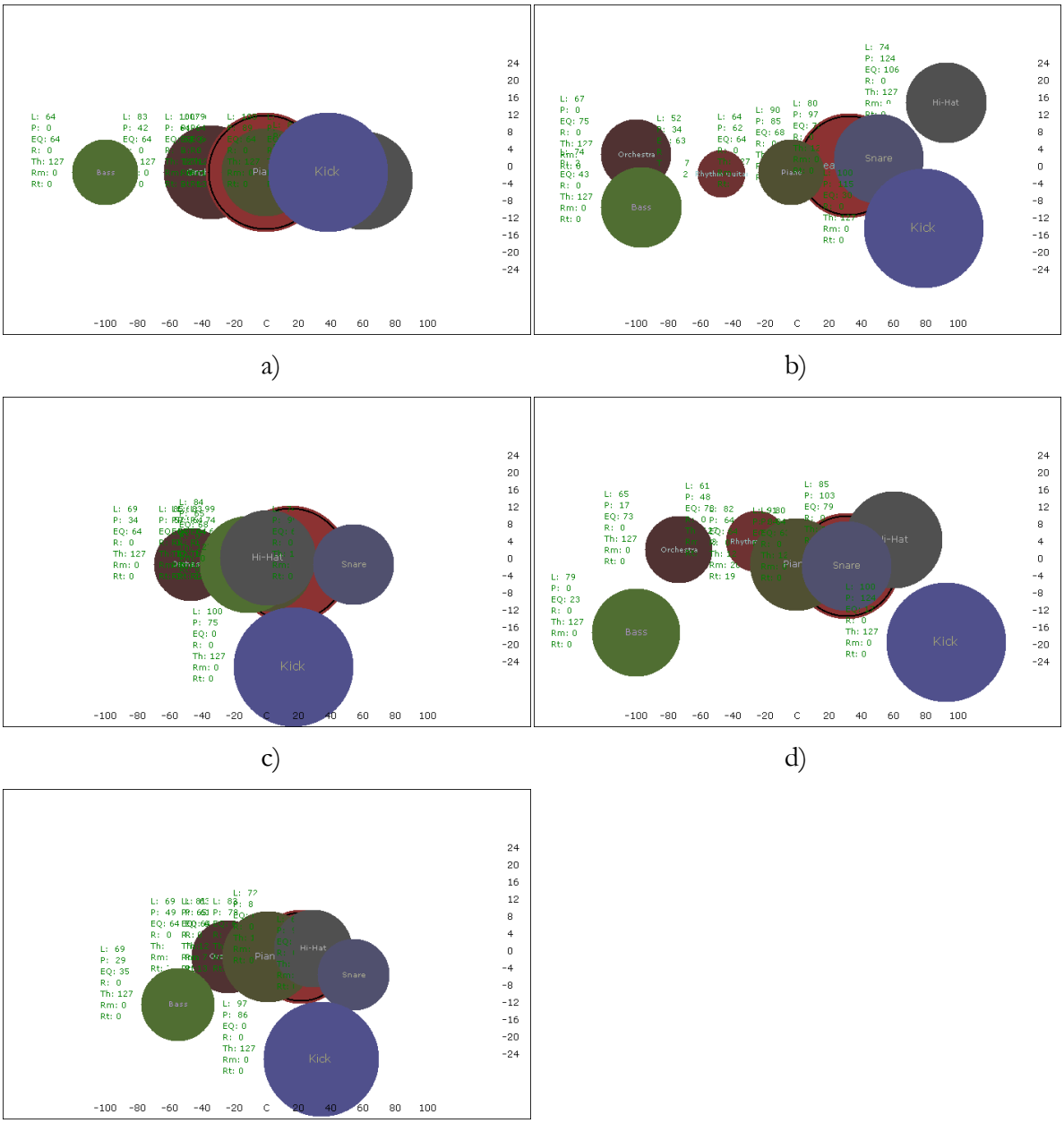
	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	64	127	0	0	0
Snare	96	57	64	127	0	0	0
H-H	75	57	64	127	0	0	0
Bass	84	64	64	127	0	0	0
Piano	92	112	64	127	0	0	0
Lead	100	64	64	107	127	5	108
Rth.	70	12	64	127	0	12	8
Orch.	86	64	64	127	0	0	0

d)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	1	127	0	0	0
Snare	91	49	89	127	0	0	0
H-H	89	39	55	127	0	0	0
Bass	87	64	0	127	0	0	0
Piano	88	127	64	127	0	0	0
Lead	100	64	64	84	55	13	78
Rth.	62	0	9	127	0	0	0
Orch.	80	64	64	127	0	0	0

e)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	64	127	0	0	0
Snare	96	57	64	127	0	0	0
H-H	75	57	64	127	0	0	0
Bass	84	64	64	127	0	0	0
Piano	92	112	64	127	0	0	0
Lead	100	64	64	107	127	5	108
Rth.	70	12	64	127	0	12	8
Orch.	86	64	64	127	0	0	0



Rys. C.2 Wizualizacje zgrań realizatora 2: (a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

Tabela C.2 Wartości kontrolerów parametrów zgrań realizatora 2

(a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

a)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	89	64	127	0	0	0
Snare	64	94	64	127	0	0	0
H-H	86	103	59	127	0	0	0
Bass	64	0	64	127	0	0	0
Piano	79	64	64	127	0	11	13
Lead	100	64	64	127	0	0	0
Rth.	66	64	64	127	0	5	11
Orch.	83	42	64	127	0	0	0

b)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	115	30	127	0	0	0
Snare	80	97	72	127	0	0	0
H-H	74	124	106	127	0	0	0
Bass	74	2	43	127	0	0	0
Piano	64	62	64	127	0	9	10
Lead	90	85	68	127	0	9	17
Rth.	52	34	63	127	0	12	16
Orch.	67	0	75	127	0	0	0

c)

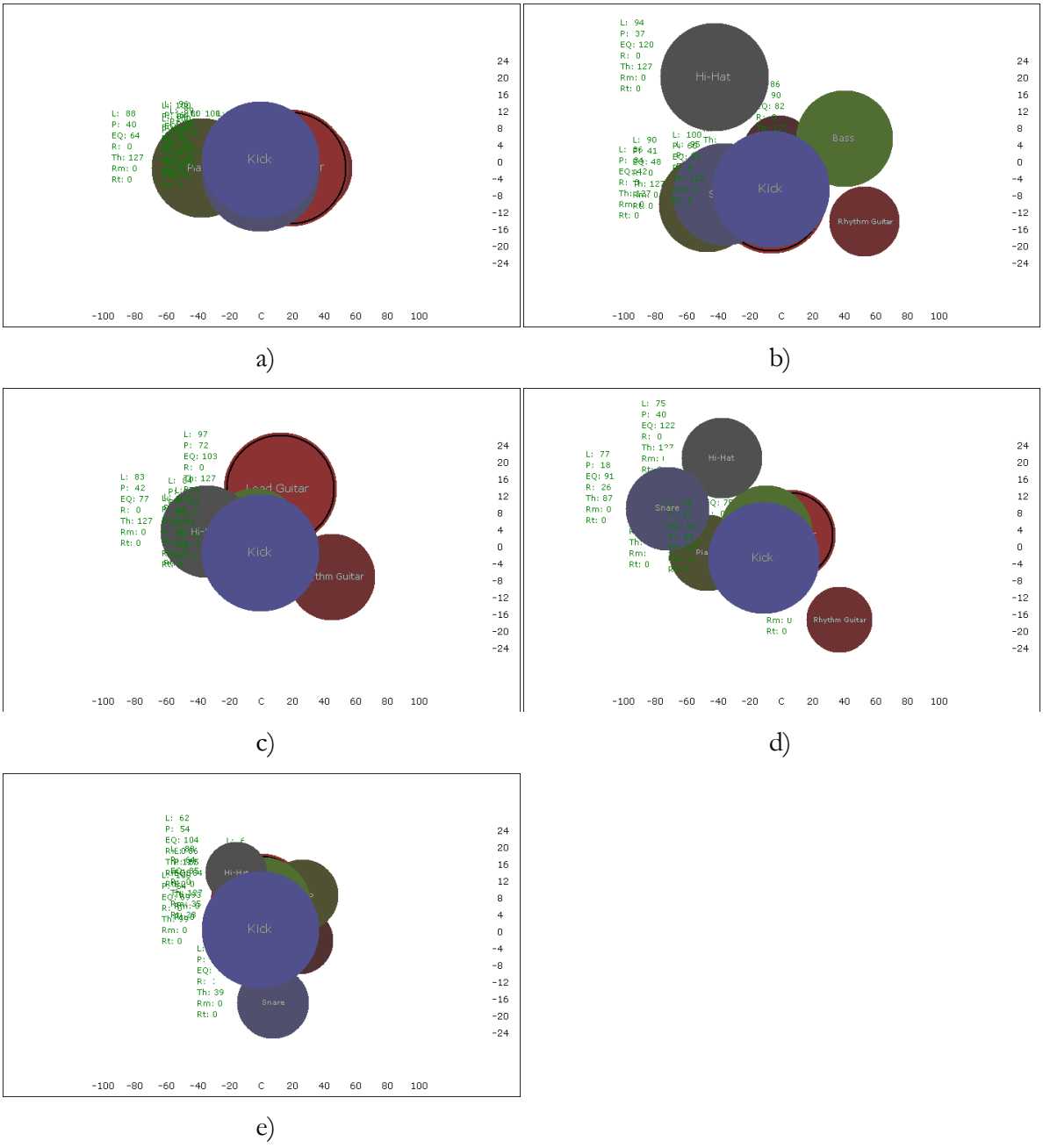
	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	75	0	127	0	0	0
Snare	74	99	64	127	0	0	0
H-H	84	65	68	127	0	0	0
Bass	85	57	64	127	0	0	0
Piano	83	64	64	127	0	9	22
Lead	99	74	64	127	0	2	1
Rth.	61	55	64	127	0	8	16
Orch.	69	34	64	127	0	0	0

d)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	124	17	127	0	0	0
Snare	80	84	63	127	0	0	0
H-H	85	103	79	127	0	0	0
Bass	79	0	23	127	0	0	0
Piano	82	64	64	127	0	20	19
Lead	91	84	63	127	0	6	12
Rth.	61	48	78	127	0	23	14
Orch.	65	17	73	127	0	0	0

e)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	97	86	0	127	0	0	0
Snare	68	99	53	127	0	0	0
H-H	72	83	69	127	0	0	0
Bass	69	29	35	127	0	0	0
Piano	81	65	64	127	0	7	13
Lead	83	78	64	127	0	6	9
Rth.	63	61	64	127	0	21	18
Orch.	69	49	64	127	0	6	7



Rys. C.3 Wizualizacje zgrań realizatora 3: (a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

Tabela C.3 Wartości kontrolerów parametrów zgrań realizatora 3

(a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

a)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	69	127	0	0	0
Snare	100	64	61	127	0	0	0
H-H	89	64	66	127	0	0	0
Bass	96	64	70	127	0	0	0
Piano	88	40	64	127	0	0	0
Lead	100	76	64	127	0	12	29
Rth.	85	82	64	127	0	0	0
Orch.	80	64	64	127	0	29	25

b)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	60	51	127	0	0	0
Snare	90	41	48	127	0	0	0
H-H	94	37	120	127	0	0	0
Bass	86	90	82	127	0	0	0
Piano	86	34	42	127	0	0	0
Lead	95	60	45	127	0	25	47
Rth.	68	98	31	127	0	0	0
Orch.	68	63	75	127	0	27	52

c)

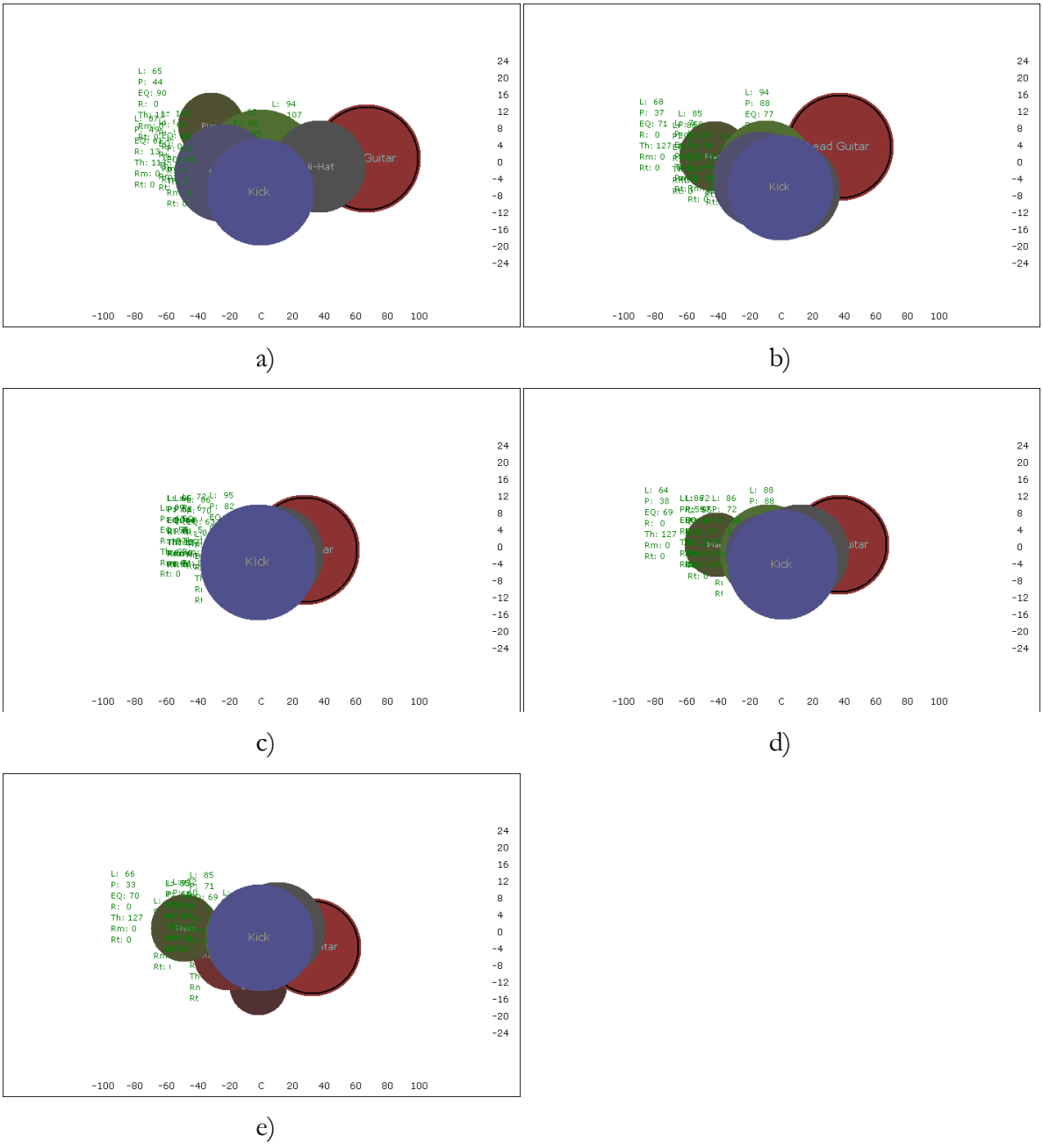
	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	64	74	22	0	0
Snare	79	63	69	127	0	0	0
H-H	83	42	77	127	0	0	0
Bass	84	62	75	118	12	0	0
Piano	84	60	65	127	0	0	0
Lead	97	72	103	127	0	13	40
Rth.	79	93	49	127	0	0	0
Orch.	77	64	64	127	0	32	28

d)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	96	57	61	50	68	0	0
Snare	77	18	91	87	26	0	0
H-H	75	40	122	127	0	0	0
Bass	83	58	77	127	0	0	0
Piano	72	34	64	127	0	0	0
Lead	82	68	75	127	0	23	25
Rth.	65	88	23	127	0	0	0
Orch.	63	61	60	127	0	40	60

e)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	69	127	0	0	0
Snare	69	69	24	39	39	0	0
H-H	62	54	104	127	0	0	0
Bass	86	65	84	127	0	0	0
Piano	69	81	90	127	0	0	0
Lead	88	64	85	127	0	35	28
Rth.	61	73	58	127	0	0	0
Orch.	66	80	62	52	6	33	57



a) b) c) d) e)

Rys. C.4 Wizualizacje zgrań realizatora 4: (a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

Tabela C.4 Wartości kontrolerów parametrów zgrań realizatora 4

(a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

a)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	93	64	49	3	4	0	0
Snare	87	49	61	117	13	0	0
H-H	83	88	65	127	0	0	0
Bass	100	64	64	127	0	0	0
Piano	65	44	90	127	0	0	0
Lead	94	107	70	127	0	22	18
Rth.	67	53	59	127	0	0	0
Orch.	56	64	51	127	0	7	4

b)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	92	64	52	43	48	0	0
Snare	86	56	57	81	20	0	0
H-H	84	69	50	127	0	0	0
Bass	85	58	64	127	0	0	0
Piano	68	37	71	127	0	0	0
Lead	94	88	77	127	0	37	64
Rth.	70	52	58	127	0	0	0
Orch.	69	64	55	127	0	46	74

c)

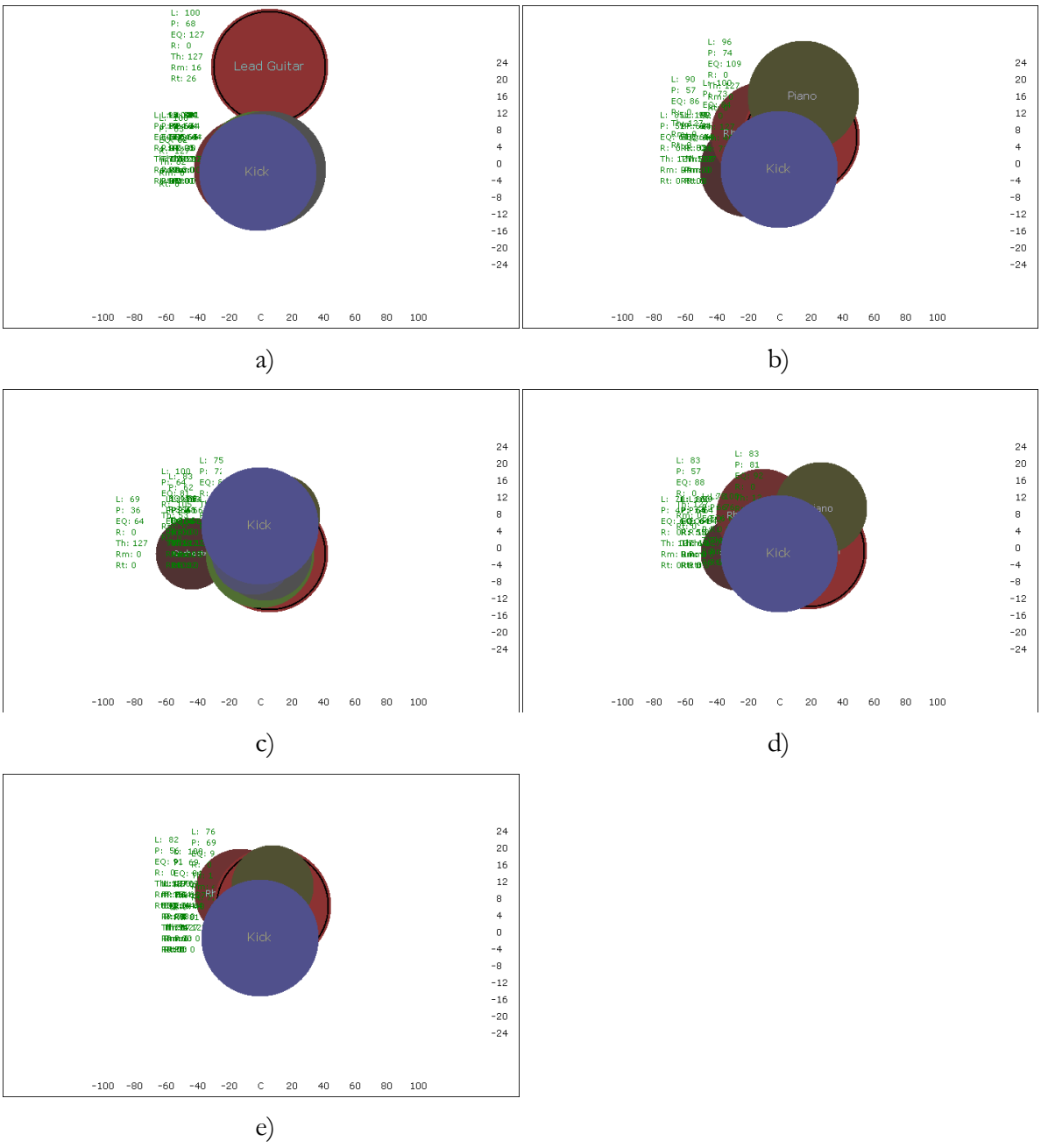
	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	99	63	58	93	9	0	0
Snare	72	64	65	118	5	0	0
H-H	86	70	63	127	0	0	0
Bass	85	62	64	127	0	0	0
Piano	66	56	64	127	0	0	0
Lead	95	82	66	127	0	32	52
Rth.	63	58	64	127	0	0	0
Orch.	55	64	42	127	0	98	46

d)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	95	65	57	63	55	0	0
Snare	72	57	64	84	11	0	0
H-H	86	72	64	127	0	0	0
Bass	86	59	64	127	0	0	0
Piano	64	38	69	127	0	0	0
Lead	88	88	69	127	0	22	67
Rth.	63	54	64	127	0	0	0
Orch.	55	64	46	127	0	68	80

e)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	93	64	64	64	54	0	0
Snare	72	60	65	85	18	0	0
H-H	85	71	69	127	0	0	0
Bass	85	61	64	127	0	0	0
Piano	66	33	70	127	0	0	0
Lead	87	85	58	127	0	25	63
Rth.	68	51	53	127	0	0	0
Orch.	59	63	34	127	0	61	51



Rys. C.5 Wizualizacje zgrań realizatora 5: (a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

Tabela C.5 Wartości kontrolerów parametrów zgrań realizatora 5

(a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

a)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	63	62	62	127	0	0
Snare	89	64	64	70	25	0	0
H-H	100	67	64	127	0	0	0
Bass	100	64	64	127	0	0	0
Piano	81	64	64	127	0	0	0
Lead	100	68	127	127	0	16	26
Rth.	90	58	64	127	0	0	0
Orch.	84	64	64	127	0	0	0

b)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	64	52	53	0	0
Snare	92	64	64	88	36	0	0
H-H	100	64	64	127	0	0	0
Bass	95	64	64	127	0	0	0
Piano	96	74	109	127	0	0	0
Lead	100	73	84	127	0	40	73
Rth.	90	57	86	127	0	0	0
Orch.	85	51	64	127	0	0	0

c)

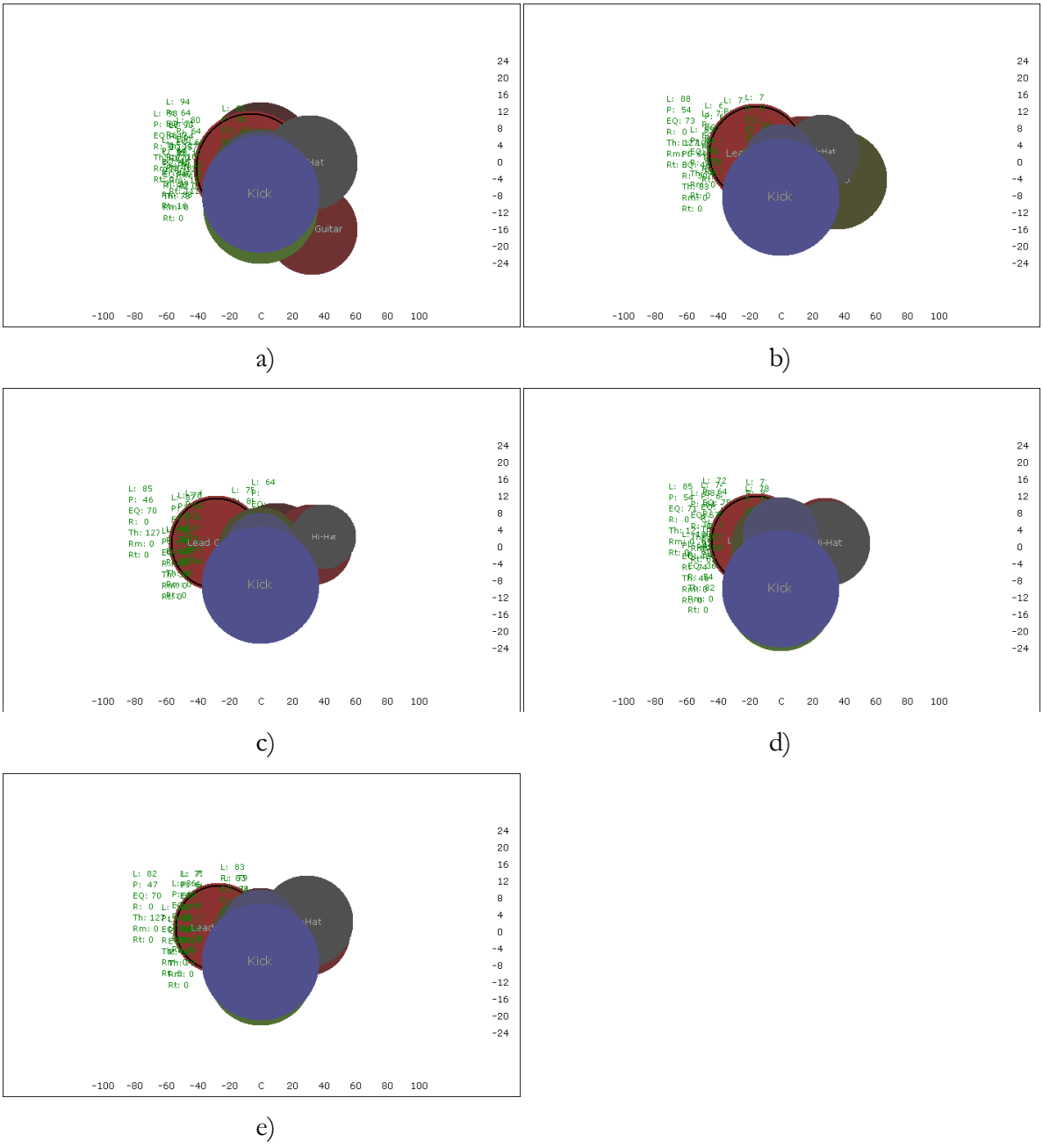
	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	81	53	105	0	0
Snare	76	61	64	127	0	0	0
H-H	84	66	64	127	0	0	0
Bass	94	64	64	127	0	0	0
Piano	75	72	88	127	0	3	4
Lead	100	68	64	127	0	28	32
Rth.	83	62	78	127	0	0	0
Orch.	69	36	64	127	0	0	0

d)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	64	64	55	0	0
Snare	86	60	64	104	18	0	0
H-H	78	66	66	127	0	0	0
Bass	89	64	64	127	0	0	0
Piano	83	81	92	127	0	43	71
Lead	100	76	66	127	0	38	42
Rth.	83	57	88	127	0	0	0
Orch.	70	47	64	127	0	0	0

e)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	64	54	71	0	0
Snare	87	61	64	87	48	0	0
H-H	89	66	64	127	0	0	0
Bass	93	64	64	127	0	0	0
Piano	76	69	96	127	0	4	72
Lead	100	69	84	127	0	40	81
Rth.	82	56	91	127	0	0	0
Orch.	73	57	64	127	0	0	0



a) b)

c) d)

e)

Rys. C.6 Wizualizacje zgrań realizatora 6: (a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

Tabela C.6 Wartości kontrolerów parametrów zgrań realizatora 6

(a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

a)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	48	29	40	38	16
Snare	80	64	60	102	12	0	0
H-H	85	84	67	127	0	0	0
Bass	98	64	40	78	22	0	0
Piano	90	64	57	127	0	15	111
Lead	98	60	64	127	0	0	0
Rth.	82	85	26	127	0	0	0
Orch.	94	64	71	127	0	109	48

b)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	46	63	30	0	0
Snare	69	64	69	82	24	0	0
H-H	71	81	74	127	0	0	0
Bass	89	64	54	96	100	0	0
Piano	88	87	57	127	0	0	78
Lead	88	54	73	127	0	0	0
Rth.	73	73	72	127	0	0	0
Orch.	73	64	64	127	0	60	82

c)

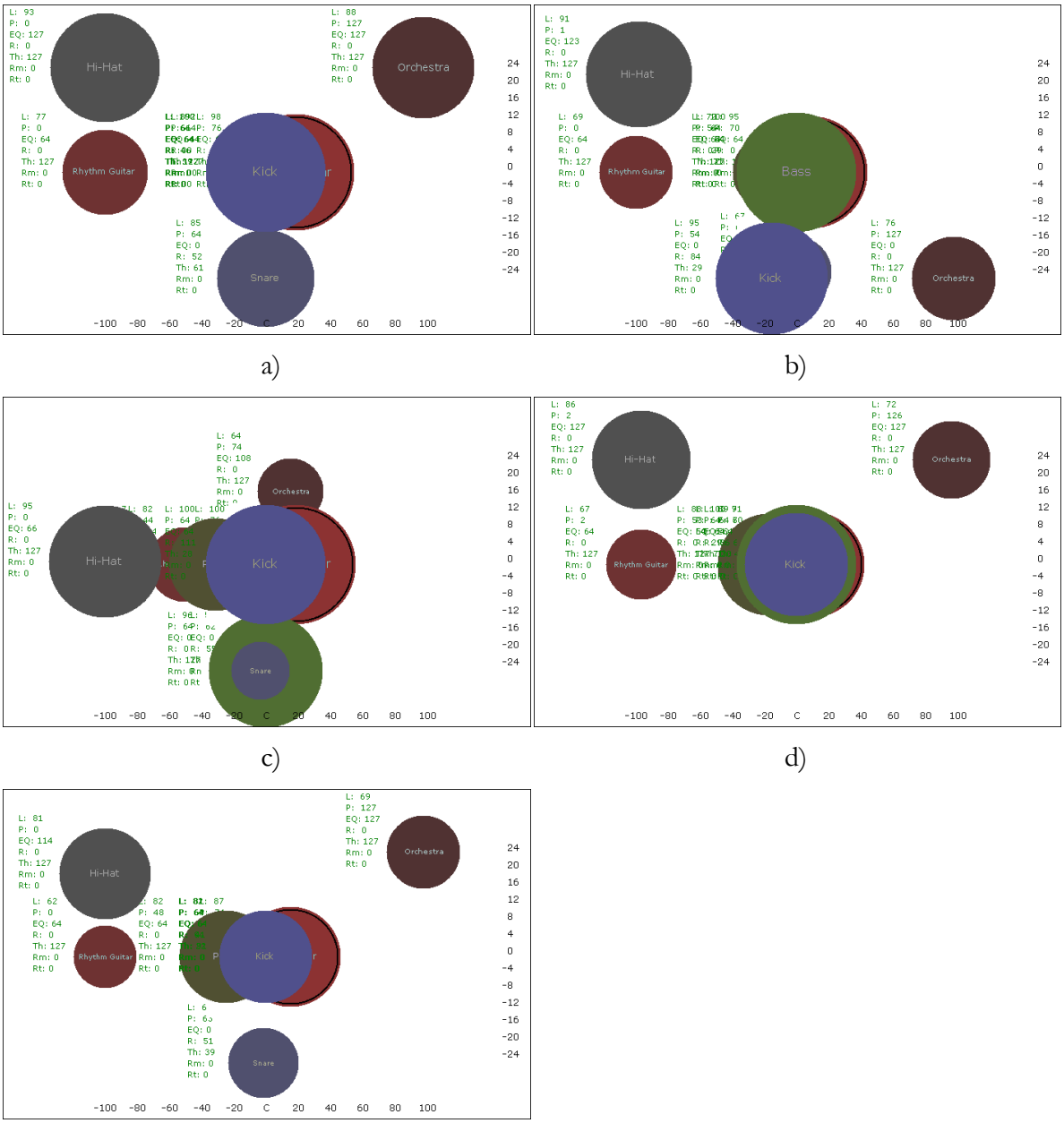
	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	44	33	48	0	0
Snare	68	64	67	79	41	0	0
H-H	64	90	74	127	0	0	0
Bass	94	64	45	88	56	0	0
Piano	78	64	66	127	0	35	44
Lead	85	46	70	127	0	0	0
Rth.	75	85	69	127	0	0	0
Orch.	87	64	64	127	0	84	31

d)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	42	46	74	0	0
Snare	72	64	75	83	69	0	0
H-H	78	83	70	127	0	0	0
Bass	92	64	36	82	54	0	0
Piano	88	64	67	127	0	36	82
Lead	85	54	71	127	0	0	0
Rth.	73	82	74	127	0	0	0
Orch.	74	64	72	127	0	61	83

e)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	49	44	88	0	0
Snare	73	64	70	83	59	0	0
H-H	83	83	74	127	0	0	0
Bass	91	64	42	73	65	0	0
Piano	86	64	64	127	0	38	86
Lead	82	47	70	127	0	0	0
Rth.	79	83	67	127	0	0	0
Orch.	75	64	70	127	0	76	88



Rys. C.7 Wizualizacje zgrań realizatora 7: (a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

Tabela C.7 Wartości kontrolerów parametrów zgrań realizatora 7

(a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

a)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	64	39	46	0	0
Snare	85	64	0	61	52	0	0
H-H	93	0	127	127	0	0	0
Bass	92	64	64	127	0	0	0
Piano	89	61	64	127	0	0	0
Lead	98	76	64	127	0	0	0
Rth.	77	0	64	127	0	0	0
Orch.	88	127	127	127	0	0	0

b)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	95	54	0	29	84	0	0
Snare	67	64	4	31	89	0	0
H-H	91	1	123	127	0	0	0
Bass	100	64	64	71	29	0	0
Piano	72	54	64	127	0	0	0
Lead	95	70	64	127	0	0	0
Rth.	69	0	64	127	0	0	0
Orch.	76	127	0	127	0	0	0

c)

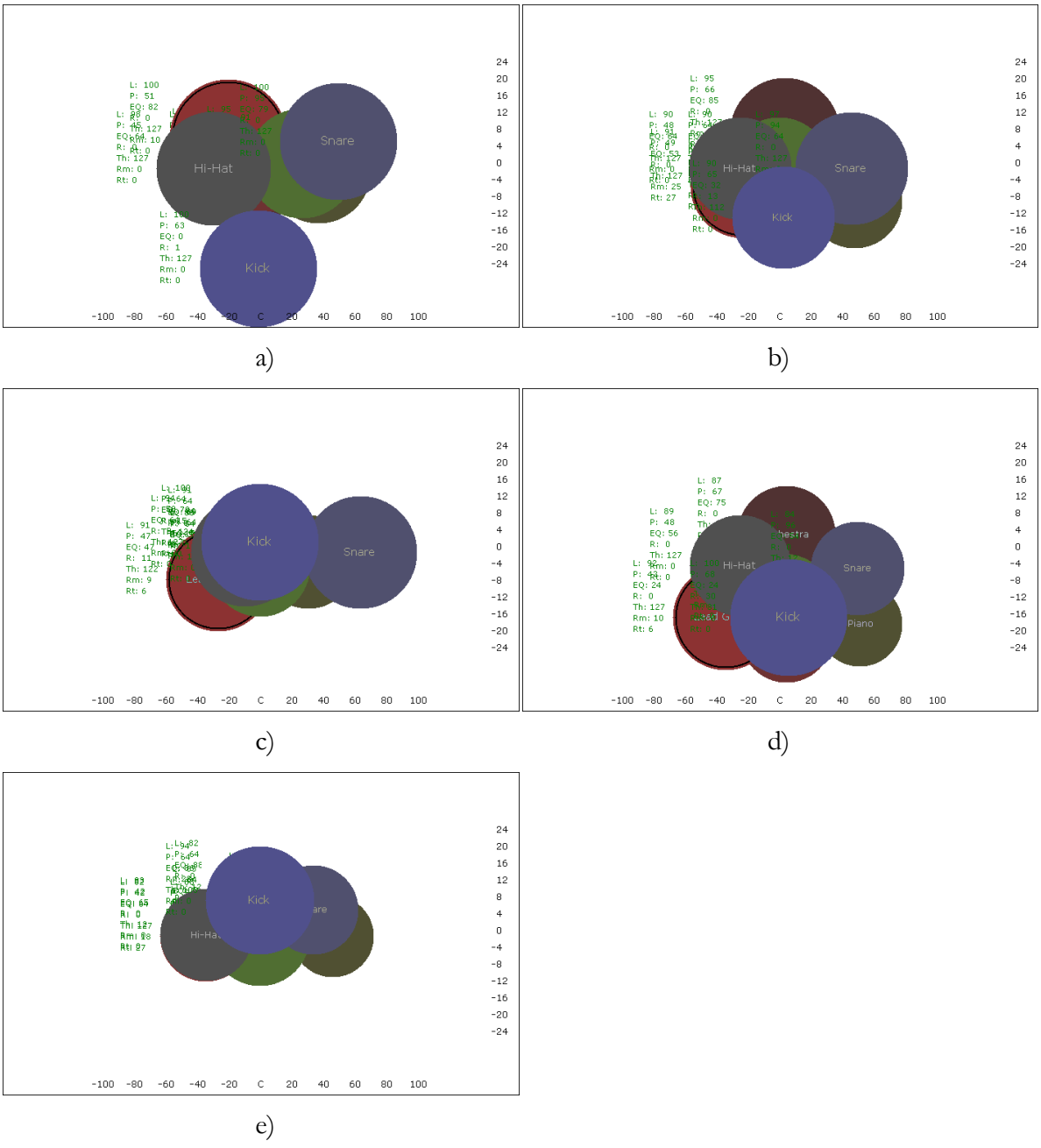
	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	64	28	111	0	0
Snare	59	62	0	10	55	0	0
H-H	95	0	66	127	0	0	0
Bass	96	64	0	127	0	0	0
Piano	82	44	64	127	0	0	0
Lead	100	76	64	127	0	0	0
Rth.	70	32	64	127	0	0	0
Orch.	64	74	108	127	0	0	0

d)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	89	64	64	30	36	0	0
Snare	71	64	64	40	61	0	0
H-H	86	2	127	127	0	0	0
Bass	100	64	64	71	29	0	0
Piano	88	53	64	127	0	0	0
Lead	91	70	64	127	0	0	0
Rth.	67	2	64	127	0	0	0
Orch.	72	126	127	127	0	0	0

e)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	82	64	64	22	94	0	0
Snare	67	63	0	39	51	0	0
H-H	81	0	114	127	0	0	0
Bass	81	64	64	91	41	0	0
Piano	82	48	64	127	0	0	0
Lead	87	74	64	127	0	0	0
Rth.	62	0	64	127	0	0	0
Orch.	69	127	127	127	0	0	0



Rys. C.8 Wizualizacje zgrań realizatora 8: (a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

Tabela C.8 Wartości kontrolerów parametrów zgrań realizatora 8

(a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

a)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	63	0	127	1	0	0
Snare	100	95	75	127	0	0	0
H-H	98	45	64	127	0	0	0
Bass	95	81	67	118	20	0	0
Piano	91	88	62	117	35	0	0
Lead	100	51	82	127	0	10	0
Rth.	85	64	66	127	0	8	14
Orch.	89	64	64	127	0	18	10

b)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	90	65	32	112	13	0	0
Snare	97	94	64	127	0	0	0
H-H	90	48	64	127	0	0	0
Bass	90	64	64	127	0	0	0
Piano	85	95	44	103	10	0	0
Lead	91	49	53	127	0	25	27
Rth.	90	64	48	127	0	30	29
Orch.	95	66	85	127	0	24	66

c)

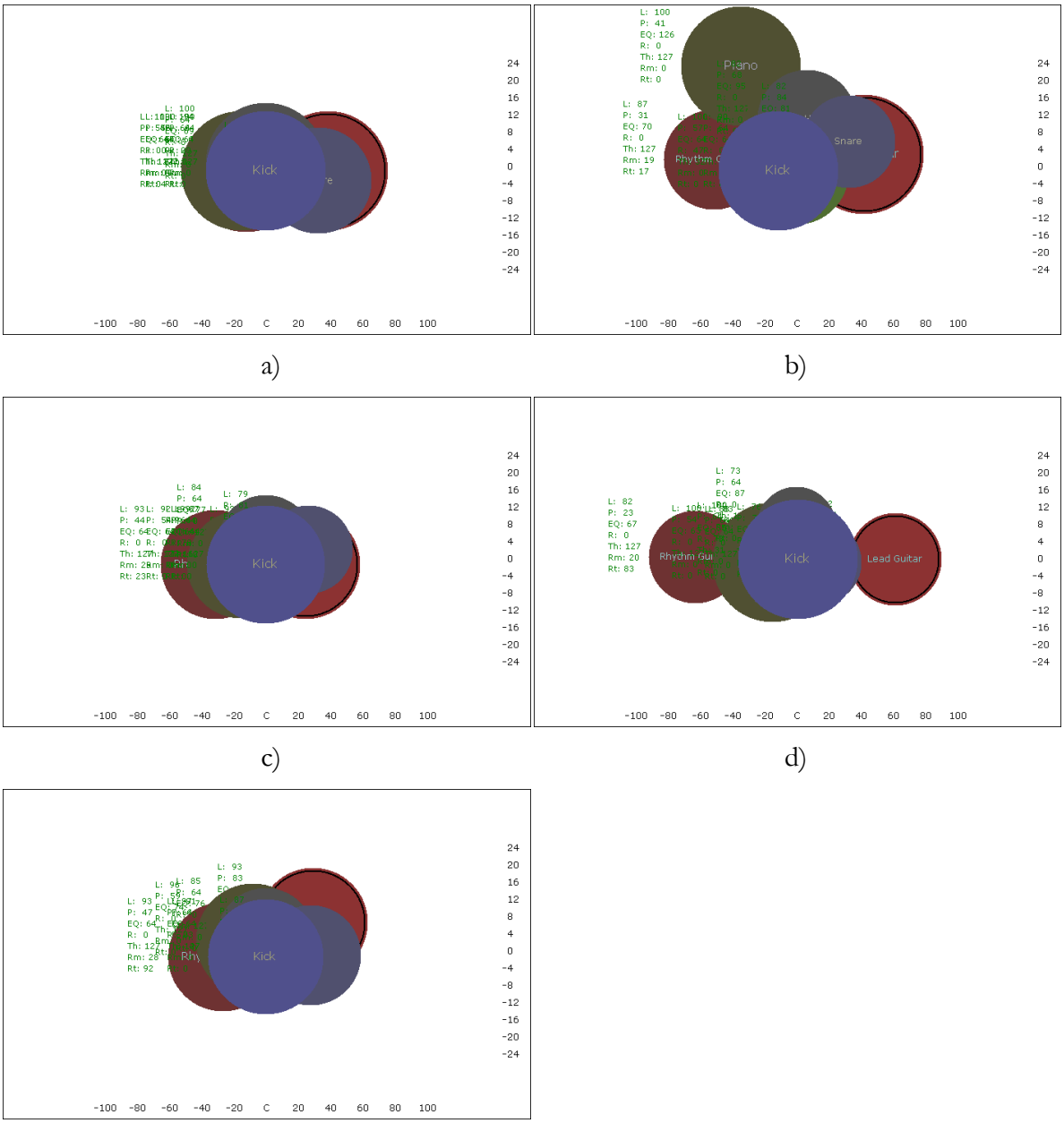
	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	70	124	15	0	0
Snare	97	105	64	127	0	0	0
H-H	94	58	64	127	0	0	0
Bass	88	64	55	116	16	0	0
Piano	84	84	58	108	14	0	0
Lead	91	47	47	122	11	9	6
Rth.	86	64	56	127	0	18	13
Orch.	91	64	69	127	0	0	0

d)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	68	24	81	30	0	0
Snare	84	96	54	127	0	0	0
H-H	89	48	56	127	0	0	0
Bass	89	66	32	94	29	0	0
Piano	78	97	20	105	8	0	0
Lead	92	42	24	127	0	10	6
Rth.	83	67	12	127	0	22	32
Orch.	87	67	75	127	0	28	64

e)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	94	64	86	103	28	0	0
Snare	81	86	80	127	0	0	0
H-H	83	42	65	127	0	0	0
Bass	88	64	64	103	52	0	0
Piano	76	94	64	127	0	0	0
Lead	82	42	64	127	0	18	27
Rth.	82	64	88	127	0	27	33
Orch.	85	64	72	127	0	29	69



a) b)

c) d)

e)

Rys. C.9 Wizualizacje zgrań realizatora 9: (a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

Tabela C.9 Wartości kontrolerów parametrów zgrań realizatora 9

(a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

a)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	64	0	20	0	0
Snare	91	85	59	127	0	0	0
H-H	100	64	69	127	0	0	0
Bass	100	64	64	127	0	0	0
Piano	100	54	64	127	0	0	0
Lead	100	89	64	127	0	0	0
Rth.	100	56	64	127	0	9	4
Orch.	94	64	64	127	0	0	0

b)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	57	64	32	47	0	0
Snare	82	84	81	127	0	0	0
H-H	82	68	95	127	0	0	0
Bass	90	64	64	127	0	0	0
Piano	100	41	126	127	0	0	0
Lead	100	90	73	127	0	0	0
Rth.	87	31	70	127	0	19	17
Orch.	81	64	64	127	0	0	0

c)

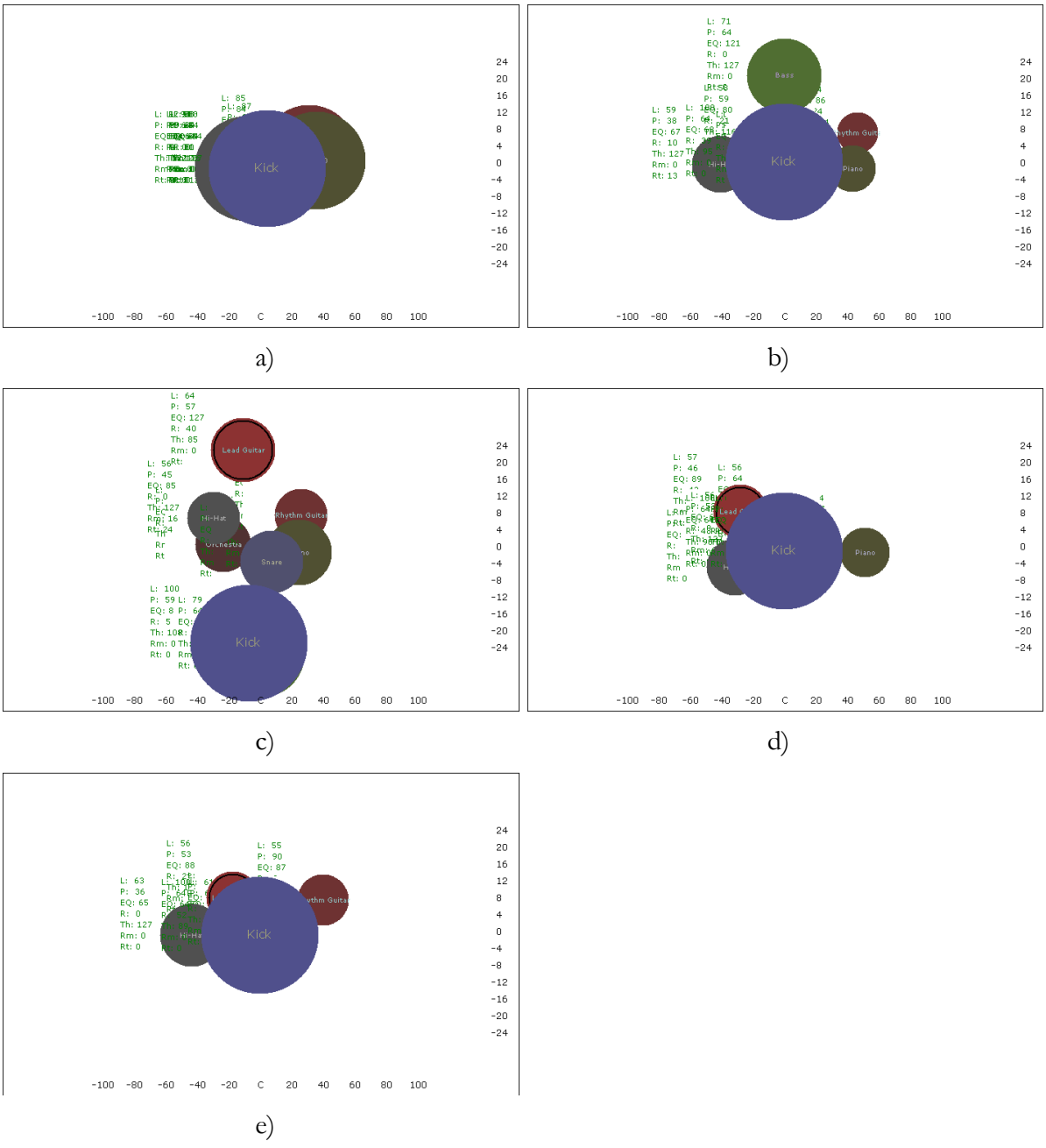
	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	99	64	64	16	72	0	0
Snare	79	81	73	127	0	0	0
H-H	84	64	77	127	0	0	0
Bass	92	64	64	127	0	0	0
Piano	92	54	64	127	0	0	0
Lead	93	80	64	127	0	0	0
Rth.	93	44	64	127	0	28	23
Orch.	87	64	64	127	0	0	0

d)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	65	31	73	0	0
Snare	76	73	64	127	0	0	0
H-H	73	64	87	127	0	0	0
Bass	88	64	64	127	0	0	0
Piano	100	54	63	127	0	0	0
Lead	82	103	66	127	0	0	0
Rth.	82	23	67	127	0	20	83
Orch.	83	64	64	127	0	0	0

e)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	97	64	64	14	73	0	0
Snare	87	82	65	127	0	0	0
H-H	85	64	76	127	0	0	0
Bass	97	64	64	127	0	0	0
Piano	96	59	74	127	0	0	0
Lead	93	83	85	127	0	0	0
Rth.	93	47	64	127	0	28	92
Orch.	91	64	64	127	0	0	0



Rys. C.10 Wizualizacje zgrań realizatora 9: (a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

Tabela C.10 Wartości kontrolerów parametrów zgrań realizatora 10

(a) obsługa za pomocą gestów / ograniczone GUI, (b) obsługa za pomocą gestów / pełne GUI, (c) obsługa za pomocą myszy i klawiatury / ograniczone GUI, (d) obsługa za pomocą myszy i klawiatury / pełne GUI, (e) bezpośrednia obsługa systemu DAW

a)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	67	64	116	51	0	0
Snare	88	64	64	127	0	0	0
H-H	92	59	64	127	0	0	0
Bass	90	64	64	127	0	0	0
Piano	87	87	69	127	0	30	127
Lead	93	64	64	127	0	0	0
Rth.	85	84	74	127	0	5	16
Orch.	83	64	64	127	0	127	127

b)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	68	95	39	0	0
Snare	59	64	64	127	0	0	0
H-H	59	38	67	127	10	0	0
Bass	71	64	121	127	0	0	0
Piano	52	92	64	127	0	39	62
Lead	58	59	80	116	21	0	0
Rth.	48	94	86	104	24	24	47
Orch.	49	64	64	127	0	49	64

c)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	59	8	108	5	0	0
Snare	63	69	58	127	0	0	0
H-H	56	45	85	127	0	0	0
Bass	79	64	0	88	20	0	0
Piano	65	80	64	103	44	48	83
Lead	64	57	127	85	40	0	0
Rth.	56	81	87	82	49	15	35
Orch.	58	49	69	127	0	72	19

d)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	64	90	48	0	0
Snare	66	64	64	127	0	0	0
H-H	59	44	55	127	0	0	0
Bass	60	64	64	127	0	0	0
Piano	54	97	64	127	0	57	67
Lead	57	46	89	84	42	0	0
Rth.	56	64	83	127	0	54	18
Orch.	56	53	66	127	0	68	69

e)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	64	89	52	0	0
Snare	64	64	68	127	0	0	0
H-H	63	36	65	127	0	0	0
Bass	65	64	64	127	0	0	0
Piano	47	80	64	127	0	34	60
Lead	56	53	88	105	28	0	0
Rth.	55	90	87	127	0	54	17
Orch.	61	64	64	127	0	75	69

Dodatek D. Analiza statystyczna rozkładu wartości parametrów

Tabela D.1 Wartości prawdopodobieństw testowych testu rang Friedmana dla poszczególnych parametrów i miar

Poziom	Zróżnicowanie Panoramy	Zróżnicowanie wzmocnienia	Próg kompresji	Stopień kompresji	Miks pogłosu	Czas pogłosu
0,000	**0,000**	**0,000**	0,664	**0,023**	**0,009**	**0,027**

Tabela D.2 Zestawienie wartości prawdopodobieństw testowych testu par rangowanych znaków Wilcoxona dla parametru *poziom*

	Gesty / tryb ograniczony	Gesty / tryb pełny	Mysz / tryb ograniczony	Mysz / tryb pełny	DAW
Gesty / tryb ograniczony		**0,000**	**0,000**	**0,000**	**0,000**
Gesty / tryb pełny			0,740	**0,007**	**0,005**
Mysz / tryb ograniczony				**0,003**	**0,003**
Mysz / tryb pełny					0,488
DAW					

Tabela D.3 Zestawienie wartości prawdopodobieństw testowych testu par rangowanych znaków Wilcoxona dla miary stopnia zróżnicowania panoramy

	Gesty / tryb ograniczony	Gesty / tryb pełny	Mysz / tryb ograniczony	Mysz / tryb pełny	DAW
Gesty / tryb ograniczony		**0,000**	0,938	**0,002**	0,49
Gesty / tryb pełny			**0,000**	0,620	**0,000**
Mysz / tryb ograniczony				**0,001**	0,302
Mysz / tryb pełny					**0,000**
DAW					

Tabela D.4 Zestawienie wartości prawdopodobieństw testowych testu par rangowanych znaków Wilcoxona dla miary stopnia zróżnicowania wzmocnienia korektora

	Gesty / tryb ograniczony	Gesty / tryb pełny	Mysz / tryb ograniczony	Mysz / tryb pełny	DAW
Gesty / tryb ograniczony		**0,000**	0,153	**0,000**	0,051
	Gesty / tryb pełny		**0,001**	0,512	**0,003**
		Mysz / tryb ograniczony		**0,007**	0,246
			Mysz / tryb pełny		0,078
				DAW	

Tabela D.5 Zestawienie wartości prawdopodobieństw testowych testu par rangowanych znaków Wilcoxona dla parametru *stopień kompresji*

	Gesty / tryb ograniczony	Gesty / tryb pełny	Mysz / tryb ograniczony	Mysz / tryb pełny	DAW
Gesty / tryb ograniczony		0,151	0,216	0,277	**0,017**
	Gesty / tryb pełny		0,804	0,808	**0,048**
		Mysz / tryb ograniczony		0,389	0,058
			Mysz / tryb pełny		0,068
				DAW	

Tabela D.6 Zestawienie wartości prawdopodobieństw testowych testu par rangowanych znaków Wilcoxona dla parametru *miks pogłosu*

	Gesty / tryb ograniczony	Gesty / tryb pełny	Mysz / tryb ograniczony	Mysz / tryb pełny	DAW
Gesty / tryb ograniczony		**0,011**	0,365	**0,035**	**0,020**
	Gesty / tryb pełny		0,289	0,313	0,244
		Mysz / tryb ograniczony		0,177	0,173
			Mysz / tryb pełny		0,804
				DAW	

Tabela D.7 Zestawienie wartości prawdopodobieństw testowych testu par rangowanych znaków Wilcoxona dla parametru *czas pogłosu*

	Gesty / tryb ograniczony	Gesty / tryb pełny	Mysz / tryb ograniczony	Mysz / tryb pełny	DAW
Gesty / tryb ograniczony		0,112	0,360	0,170	0,163
	Gesty / tryb pełny		**0,019**	0,522	0,705
		Mysz / tryb ograniczony		0,130	**0,034**
			Mysz / tryb pełny		0,193
				DAW	

Dodatek E. Analiza statystyczna przydzielonych przez ekspertów ocen walorów estetycznych zgrań

Tabela E.1 Zestawienie wartości testu rang Friedmana dla ocen przydzielonych przez ekspertów zgraniom realizatora 3

	SS Efekt	df Efekt	MS Efekt	SS Błąd	df Błąd	MS Błąd	χ^2	p
ocena zgrania	83	4	20,75	28	44	0,6364	35,89	$3{,}05 \cdot 10^{-7}$

Tabela E.2 Zestawienie wartości testu rang Friedmana dla ocen przydzielonych przez ekspertów zgraniom realizatora 7

	SS Efekt	df Efekt	MS Efekt	SS Błąd	df Błąd	MS Błąd	χ^2	P
ocena zgrania	49,4583	4	12,3646	67,5417	44	1,535	20,29	0,0004

Tabela E.3 Zestawienie wartości prawdopodobieństw testowych testu par rangowanych znaków Wilcoxona dla ocen przydzielonych przez ekspertów zgraniom realizatora 3

	Gesty / tryb ograniczony	Gesty / tryb pełny	Mysz / tryb ograniczony	Mysz / tryb pełny	DAW
Gesty / tryb ograniczony		**0,0244**	0,2031	**0,0015**	**$9{,}77 \cdot 10^{-4}$**
	Gesty / tryb pełny		0,0547	**0,0078**	**$4{,}88 \cdot 10^{-4}$**
		Mysz / tryb ograniczony		**0,002**	**$4{,}88 \cdot 10^{-4}$**
			Mysz / tryb pełny		**0,0186**
				DAW	

Tabela E.4 Zestawienie wartości testu par rangowanych znaków Wilcoxona dla ocen przydzielonych przez ekspertów zgraniom realizatora 7

	Gesty / tryb ograniczony	Gesty / tryb pełny	Mysz / tryb ograniczony	Mysz / tryb pełny	DAW
Gesty / tryb ograniczony		0,2783	**0,0068**	0,2402	0,9658
	Gesty / tryb pełny		**0,0034**	0,3013	0,0674
		Mysz / tryb ograniczony		**0,0015**	**0,0024**
			Mysz / tryb pełny		0,1055
				DAW	

Dodatek F. Analiza statystyczna wartości skuteczności rozpoznawania gestów dynamicznych

Tabela F.1 Zestawienie wartości testu rang Friedmana badania istotności statystycznej wpływu zastosowania logiki rozmytej na zwiększenie skuteczności rozpoznawania gestów dynamicznych lewej ręki

	SS Efekt	df Efekt	MS Efekt	SS Błąd	df Błąd	MS Błąd	χ^2	p
ocena zgrania	156,706	7	22,3866	277,294	112	2,4758	42,97	**0,000**

Tabela F.2 Zestawienie wartości testu rang Friedmana badania istotności statystycznej wpływu zastosowania logiki rozmytej na zwiększenie skuteczności rozpoznawania gestów dynamicznych prawej ręki

	SS Efekt	df Efekt	MS Efekt	SS Błąd	df Błąd	MS Błąd	χ^2	p
ocena zgrania	239,265	7	34,1807	220,235	112	1,9664	61,96	**0,000**

Tabela F.3 Zestawienie wartości testu rang Friedmana badania istotności statystycznej wpływu zastosowania logiki rozmytej na zwiększenie skuteczności rozpoznawania gestów dynamicznych obu rąk

	SS Efekt	df Efekt	MS Efekt	SS Błąd	df Błąd	MS Błąd	χ^2	p
ocena zgrania	14,647	7	2,0924	238,853	112	2,1326	6,88	0,4419

Tabela F.4 Zestawienie wartości prawdopodobieństw testowych testu par rangowanych znaków Wilcoxona dla wartości skuteczności rozpoznawania gestów lewej ręki, z wykorzystaniem logiki rozmytej i bez logiki rozmytej

		Bez logiki rozmytej			
		lewo	prawo	góra	Dół
Z logiką rozmytą	lewo	**0,0003**			
	prawo		**0,0003**		
	góra			**0,0001**	
	dół				**0,0002**

Tabela F.5 Zestawienie wartości prawdopodobieństw testowych testu par rangowanych znaków Wilcoxona dla wartości skuteczności rozpoznawania gestów prawej ręki, z wykorzystaniem logiki rozmytej i bez logiki rozmytej

		Bez logiki rozmytej			
		lewo	prawo	góra	Dół
Z logiką rozmytą	lewo	**0,0007**			
	prawo		**0,0007**		
	góra			0,5000	
	dół				0,3125

Dodatek G. Dokumentacja techniczna systemu

Specyfikacja wymagań względem systemu

Historia dokumentu specyfikacji wymagań

Data	Wersja	Zmiany
23.08.2010	1.0.2	Dodanie wymagań: F2, F4, F6, F7, F8, F11, F12
20.08.2010	1.0.1	Dodanie opcjonalnego wymagania (czas ataku i zwolnienia kompresji)
27.06.2010	1.0.0	Pierwsza wersja

Szablon opisu wymagań

Identyfikator:		**Priorytet:**		
Tytuł:				
Opis:				
Powiązane wymagania:				

Wymagania są opisane według jednolitego szablonu zawierającego następujące informacje:

Identyfikator

Symbol wymagania, unikalny w ramach całej specyfikacji wymagań. Sposób tworzenia identyfikatorów został szczegółowo opisany w dalszej części.

Priorytet

Ważność wymagania w odniesieniu do całego systemu. Priorytet może przyjmować trzy wartości:

1 – podstawowy – wymaganie musi być koniecznie spełnione,

2 – przydatny – wymaganie powinno być spełnione, jeśli starczy na to czasu podczas realizacji systemu,

3 – rozszerzony – wymaganie dotyczy funkcji, która może być wprowadzona w następnej wersji systemu, pokazuje prawdopodobny kierunek rozwoju systemu.

Tytuł

Aspekt systemu lub procesu jego wytwarzania, omówiony w danym wymaganiu.

Opis

Treść wymagania.

Powiązane wymagania

Identyfikatory innych wymagań, które są w jakiś sposób powiązane z danym wymaganiem.

Identyfikatory wymagań

Każde wymaganie ma swój unikalny identyfikator, który składa się z trzech części:

Typ[Grupa]Numer,

gdzie:

Typ – symbol typu wymagania

O – ogólne

F – funkcjonalne

N – niefunkcjonalne

Grupa – symbol grupy wymagań;

dla wymagań funkcjonalnych:

dla wymagań pozafunkcjonalnych:

B – bezpieczeństwo (*safety*)

D – dokumentacja

IP – interfejs graficzny pacjenta

IT – interfejs graficzny terapeuty

ŚD – środowisko docelowe

ŚDi – środowisko diagnostyczne

ŚW – środowisko wytwórcze

Wi – wiarygodność

W – wydajność

Z – zabezpieczenia (*security*),

Numer – numer wymagania w obrębie danej grupy.

Wymagania ogólne

Identyfikator:	O1	**Priorytet:**	1	
Tytuł:	Cel systemu			
Opis:	System powinien umożliwiać przeprowadzenie typowych procesów miksowania sygnałów wchodzących w skład nagrania muzycznego za pomocą gestów rąk (palców, przedramion).			
Powiązane Wymagania:				

Identyfikator:	O2	**Priorytet:**	1	
Tytuł:	Komponenty systemu			
Opis:	Komponentami systemu są komputer klasy PC lub laptop, rzutnik multimedialny, kamera internetowa podłączona do portu USB, ekran dla obrazu wyświetlanego przez rzutnik.			
Powiązane Wymagania:				

Identyfikator:	O3	**Priorytet:**	2	
Tytuł:	Liczba użytkowników			
Opis:	Procesy miksowania obsługuje jedna osoba.			
Powiązane Wymagania:				

Identyfikator:	O4	**Priorytet:**	1	
Tytuł:	Umiejscowienie użytkownika			
Opis:	Użytkownik znajduje się pomiędzy rzutnikiem a ekranem dla obrazu z rzutnika, w takiej odległości od ekranu, aby przy widocznym obrazie z rzutnika wyciągnięte przed siebie ręce użytkownika rzucały cień. Cień pozostałych części ciała użytkownika nie powinien być widoczny w wyświetlanym obrazie. Cień lewej ręki zajmuje lewą połowę ekranu, prawej ręki – prawą.			
Powiązane Wymagania:	O3			

Identyfikator:	O5	**Priorytet:**	1	
Tytuł:	Sposób odtwarzania dźwięku			
Opis:	System odtwarza dźwięk w trybie obsługiwanym przez wybrane oprogramowanie DAW.			
Powiązane Wymagania:				

Wymagania funkcjonalne

Identyfikator:	F1	**Priorytet:**	1	
Tytuł:	Wczytanie sesji			
Opis:	System powinien umożliwiać wczytanie wcześniej zapisanej sesji, tj. wszelkie elementy interfejsu przyjmują stan zgodny ze stanem zapamiętanym dla danej sesji.			
Powiązane Wymagania:	Brak			
Powiązane przypadki użycia:	PU1			

Identyfikator:	F2	**Priorytet:**	1	
Tytuł:	Ustalenie kolejności ścieżek			
Opis:	System umożliwia ustalenie kolejności ścieżek, tj. przypisanie każdej ścieżce liczby, począwszy od liczby 1.			
Powiązane Wymagania:	F1			
Powiązane przypadki użycia:	PU2			

Identyfikator:	F3	**Priorytet:**	1	
Tytuł:	Wybór ścieżek do edycji			
Opis:	System powinien umożliwiać wybór ścieżek do edycji parametrów, tj. wszelkie operacje przetwarzania dźwięku mogą być wykonywane indywidualnie dla tych ścieżek. Możliwy jest wybór dwóch ścieżek (jednej dla lewej ręki i jednej dla prawej).			
Powiązane Wymagania:	F2			
Powiązane przypadki użycia:	PU3			

Identyfikator:	F4	**Priorytet:**	1	
Tytuł:	Zaznaczenie ścieżki			
Opis:	System umożliwia zaznaczenie jednej z dwóch wybranych ścieżek, tj. wszelkie operacje przetwarzania dźwięku, jeśli użytkownik nie określi inaczej, związane są z tą ścieżką.			
Powiązane Wymagania:	F3			
Powiązane przypadki użycia:	PU4			

Identyfikator:	F5	**Priorytet:**	1	
Tytuł:	Odtwarzanie dźwięku			
Opis:	System powinien zapewniać jednoczesne odtwarzanie wszystkich niewyciszonych ścieżek.			
Powiązane Wymagania:	Brak			
Powiązane przypadki użycia:	PU5			

Identyfikator:	F6	**Priorytet:**	1	
Tytuł:	Zatrzymanie odtwarzania dźwięku			
Opis:	System umożliwia zatrzymanie odtwarzania wszystkich niewyciszonych ścieżek. Ponowne odtwarzanie dźwięku jest realizowane od momentu zatrzymania.			
Powiązane Wymagania:	F5			
Powiązane przypadki użycia:	PU6			

Identyfikator:	F7	**Priorytet:**	1	
Tytuł:	Przewijanie w przód			
Opis:	System umożliwia szybsze osiągnięcie punktu na osi czasu nagrania niż to wynika z rzeczywistego tempa z jakim odtwarzane jest nagranie. Przydatne, ale nie wymagane, jest umożliwienie przewijania z odtwarzaniem. Tempo odtwarzania odpowiada szybkości przewijania.			
Powiązane Wymagania:	F5			
Powiązane przypadki użycia:	PU7			

Identyfikator:	F8	**Priorytet:**	1	
Tytuł:	Przewijanie w tył			
Opis:	System umożliwia osiągnięcie wcześniejszego punktu na osi czasu nagrania niż aktualny. Przydatne, ale nie wymagane, jest umożliwienie przewijania z odtwarzaniem. Tempo odtwarzania odpowiada szybkości przewijania.			
Powiązane Wymagania:	F5			
Powiązane przypadki użycia:	PU8			

Identyfikator:	F9	**Priorytet:**	1	
Tytuł:	Odtwarzanie tylko wybranych ścieżek			
Opis:	System udostępnia możliwość wybrania pojedynczej ścieżki do odtwarzania (tryb SOLO), przy czym pozostałe ścieżki nie są wówczas odtwarzane. Istnieje możliwość wyboru dowolnej liczby ścieżek do odtwarzania. Ustawienie ścieżki w tryb SOLO powoduje wyciszenie wszystkich pozostałych ścieżek (poza ścieżkami również w trybie SOLO).			
Powiązane Wymagania:	F5			
Powiązane przypadki użycia:	PU9			

cia:	

Identyfikator:	F10	**Priorytet:**	1	
Tytuł:	Wyciszenie ścieżki			
Opis:	Możliwe jest wyciszenie wybranej ścieżki (tryb MUTE), tj. odtwarzane są wszystkie ścieżki poza nią. Użytkownik może wyciszyć dowolną liczbę ścieżek.			
Powiązane Wymagania:	F5			
Powiązane przypadki użycia:	PU10			

Identyfikator:	F11	**Priorytet:**	1	
Tytuł:	Dezaktywacja wszystkich trybów SOLO			
Opis:	Możliwe jest dezaktywowanie, w jednym kroku, trybów SOLO dla wszystkich ścieżek, tj. odtwarzane są znów wszystkie ścieżki (poza wyciszonymi).			
Powiązane Wymagania:	F5, F10			
Powiązane przypadki użycia:	PU11			

Identyfikator:	F12	**Priorytet:**	1	
Tytuł:	Dezaktywacja wszystkich trybów wyciszenia			
Opis:	Możliwe jest dezaktywowanie, w jednym kroku, trybów wyciszenia dla wszystkich ścieżek, tj. odtwarzane są znów wszystkie ścieżki.			
Powiązane Wymagania:	F5			
Powiązane przypadki użycia:	PU12			

Identyfikator:	F13	**Priorytet:**	1	
Tytuł:	Zmiana poziomu ścieżki			
Opis:	System umożliwia zmianę poziomu odtwarzania wybranej ścieżki, od -80 dB, oznaczającego całkowite wyciszenie, do poziomu 0dB.			
Powiązane Wymagania:	F5			
Powiązane przypadki użycia:	PU13			

Identyfikator:	F14	**Priorytet:**	1	
Tytuł:	Panoramowanie dźwięku			
Opis:	Dla wybranej ścieżki powinno być możliwe ustawienie panoramy dźwięku, tj. użytkownik ma możliwość płynnej zmiany proporcji poziomu dźwięku między kanałami systemu odsłuchowego. Domyślnie ścieżka odtwarzana jest w centrum panoramy (z centralnego głośnika).			
Powiązane Wymagania:	F5			
Powiązane przypadki użycia:	PU14			

Identyfikator:	F15	**Priorytet:**	1	
Tytuł:	Dodanie pogłosu			
Opis:	System umożliwia wzbogacenie ścieżki o pogłos, określany za pomocą parametrów: czas pogłosu, poziom dźwięku pogłosowanego.			
Powiązane Wymagania:	Brak			
Powiązane przypadki użycia:	PU15, PU19, PU20			

Identyfikator:	F16	**Priorytet:**	1	
Tytuł:	Kompresja dynamiki			
Opis:	System powinien umożliwiać zmianę dynamiki wybranej ścieżki poprzez manipulację następującymi parametrami kompresji: próg zadziałania, stopień kompresji.			
Powiązane Wymagania:	Brak			
Powiązane przypadki użycia:	PU16, PU21, PU22			

Identyfikator:	F16a	**Priorytet:**	2	
Tytuł:	Czas ataku i zwolnienia kompresji dynamiki			
Opis:	Opcjonalnie system umożliwia zmianę czasu ataku i czasu zwolnienia kompresji dynamiki.			
Powiązane Wymagania:	F16			
Powiązane przypadki użycia:				

Identyfikator:	F17	**Priorytet:**	1	
Tytuł:	Korekcja częstotliwościowa			
Opis:	System umożliwia wzmocnienie lub osłabienie pasma określonego w oprogramowaniu DAW, zgodnie z parametrami filtru zadanymi w tym oprogramowaniu. Przydatną, ale niewymaganą funkcją jest zmiana charakterystyki częstotliwościowej odtwarzania wybranej ścieżki w zakresie 20 Hz – 20kHz, z dokładnością wyboru częstotliwości co najmniej do interwału tercji i dokładnością poziomu dla danej częstotliwości co najmniej do 1 dB.			
Powiązane Wymagania:	Brak			
Powiązane przypadki uży-	PU17			

cia:	

Identyfikator:	F18	**Priorytet:**	3	
Tytuł:	Eksport do pliku dźwiękowego			
Opis:	Możliwy jest eksport sumy wszystkich ścieżek do pliku w formacie stereofonicznym WAVE PCM lub do 5 plików monofonicznych WAVE PCM odpowiadających poszczególnym kanałom systemu wielokanałowego 5.1. Hierarchia czasowa ścieżek w pliku odzwierciedla hierarchię w interfejsie graficznym użytkownika, z poziomu którego dokonano eksportu.			
Powiązane Wymagania:	Brak			
Powiązane przypadki użycia:	PU18			

Identyfikator:	F19	**Priorytet:**	3	
Tytuł:	Automatyka			
Opis:	Dla wybranej ścieżki i wybranego parametru, takiego jak np. poziom dźwięku lub stopień kompresji, system umożliwia ustalenie krzywej automatyki.			
Powiązane Wymagania:	Brak			

Identyfikator:	I1	**Priorytet:**	1	
Tytuł:	Obsługa interfejsu			
Opis:	W zakresie wykonywania funkcji związanych bezpośrednio z miksowaniem materiału muzycznego, interfejs użytkownika powinien umożliwiać zarówno sterowanie za pomocą jedynie gestów rąk (palców dłoni, przedramion) jak i za pomocą jedynie klawiatury i myszy. Dopuszcza się możliwość, aby funkcje niezwiązane bezpośrednio z procesami miksowania, takie jak np. załadowanie sesji do systemu były wykonywane jedynie za pomocą klawiatury i myszy.			
Powiązane Wymagania:	Brak			

Identyfikator:	I2	**Priorytet:**	1	
Tytuł:	Praca z podglądem / bez podglądu obrazu			
Opis:	Użytkownik powinien mieć możliwość przeprowadzenia wszystkich udostępnianych funkcji, bezpośrednio związanych z przeprowadzaniem procesów miksowania, zarówno w trybie pełnego interfejsu graficznego jak i w trybie ograniczonego interfejsu. Dopuszcza się możliwość, aby funkcje niezwiązane bezpośrednio z procesami miksowania, takie jak np. załadowanie sesji do systemu lub były wykonywane jedynie w trybie pełnego interfejsu graficznego.			
Powiązane Wymagania:	Brak			

Wymagania pozafunkcjonalne

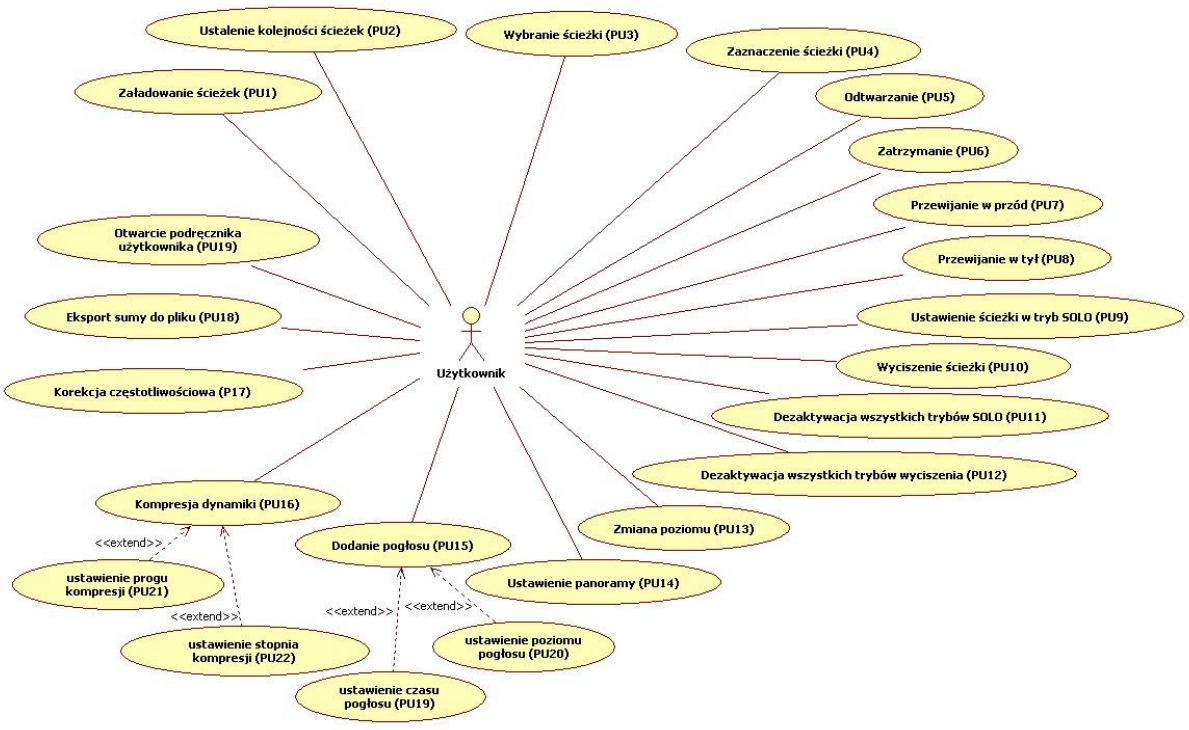
Identyfikator:	ND1	**Priorytet:**	1	
Tytuł:	Dokumentacja			
Opis:	Każdy z etapów cyklu wytwarzania będzie dokumentowany zgodnie z harmonogramem kwartalnym. Każdy z dokumentów związanych bezpośrednio z częścią wykonywanego systemu będzie opatrzony historią zmian. Dokumentacja przechowywana będzie w miejscu jej wytworzenia oraz w serwisie plone.			
Powiązane Wymagania:				

Identyfikator:	NŚD1	**Priorytet:**	1	
Tytuł:	Systemowe środowisko docelowe			
Opis:	Wymaga się, aby system działał na następujących rodzinach systemów operacyjnych: Windows XP, Windows Vista, Windows 7.			
Powiązane Wymagania:				

Identyfikator:	NŚW1	**Priorytet:**	1	
Tytuł:	Systemowe środowisko wytwórcze			
Opis:	Wymaga się, aby system był wytwarzany w środowisku zgodnym ze środowiskami wymienionymi w wymaganiu NŚD1.			
Powiązane Wymagania:				

Identyfikator:	NŚDi1	**Priorytet:**	1	
Tytuł:	Systemowe środowisko diagnostyczne			
Opis:	Wymaga się, aby system był testowany w środowiskach systemowych zgodnych z wymienionymi w wymaganiu NŚD1.			
Powiązane Wymagania:	NŚD1			

Analiza wymagań (model systemu)



Rys. G.1 Diagram przypadków użycia

Opis przypadków użycia

Aktorem biorącym udział we wszystkich przypadkach użycia jest użytkownik systemu.

Identyfikator:	PU1	
Nazwa:	Załadowanie ścieżek	
Aktorzy:	Użytkownik	
Warunki początkowe:	uruchomiona aplikacja	
Opis realizowanej funkcji:	Wybór pliku/plików na dysku i załadowanie ich do systemu w postaci ścieżek nagrania (jeden plik = jedna ścieżka)	
Sytuacje wyjątkowe:	Próba wczytania pliku o formacie innym niż obsługiwane formaty plików dźwiękowych	
Warunki końcowe:	Załadowane ścieżki/a lub informacja o błędzie w przypadku próby wczytania nieobsługiwanego formatu	
Powiązane wymagania:	F1	

Identyfikator:	PU2	
Nazwa:	Ustalenie kolejności ścieżek	
Aktorzy:	Użytkownik	
Warunki początkowe:	Załadowane ścieżki/a	
Opis realizowanej funkcji:	Użytkownik decyduje jakie liczby przypisać załadowanym ścieżkom. Liczby nie mogą się powtarzać.	
Sytuacje wyjątkowe:	Brak	
Warunki końcowe:	Ustalona kolejność ścieżek	
Powiązane wymagania:	F2	

Identyfikator:	PU3	
Nazwa:	Wybranie ścieżki	
Aktorzy:	Użytkownik	
Warunki początkowe:	Załadowane ścieżki	
Opis realizowanej funkcji:	Przypisanie ścieżki do ręki (gesty wykonywane daną ręką przekładane są na akcje związane z wybraną ścieżką); możliwość przypisania dwóch ścieżek do rąk przy pracy jednoosobowej lub czterech ścieżek przy pracy dwuosobowej.	
Sytuacje wyjątkowe:	Brak	
Warunki końcowe:	Ścieżki przypisane do rąk.	
Powiązane wymagania:	F3	

Identyfikator:	PU4	
Nazwa:	Zaznaczenie ścieżki	
Aktorzy:	Użytkownik	
Warunki początkowe:	Wybrane ścieżki	
Opis realizowanej funkcji:	Zaznaczenie jednej z dwóch ścieżek (przy pracy jednoosobowej) lub dwóch z czterech (przy pracy dwuosobowej). Wykonanie gestu angażującego obie ręce powoduje wykonanie akcji dla zaznaczonej ścieżki.	
Sytuacje wyjątkowe:	Brak	
Warunki końcowe:	Zaznaczona ścieżka	
Powiązane wymagania:	F4	

Identyfikator:	PU5	
Nazwa:	Odtwarzanie	
Aktorzy:	Użytkownik	
Warunki początkowe:	Załadowane ścieżki	

kowe:	
Opis realizowanej funkcji:	Odtwarzanie sumy ścieżek z uwzględnieniem wszystkich parametrów.
Sytuacje wyjątkowe:	Osiągnięty koniec nagrania – zatrzymanie odtwarzania i powrót do początku nagrania.
Warunki końcowe:	Odtwarzany dźwięk z głośników
Powiązane wymagania:	F5

Identyfikator:	PU6
Nazwa:	Zatrzymanie
Aktorzy:	Użytkownik
Warunki początkowe:	Odtwarzany dźwięk
Opis realizowanej funkcji:	Zatrzymanie odtwarzania dźwięku – ponowne odtwarzanie realizowane jest od miejsca zatrzymania.
Sytuacje wyjątkowe:	Brak
Warunki końcowe:	Zatrzymane odtwarzanie
Powiązane wymagania:	F6

Identyfikator:	PU7
Nazwa:	Przewijanie w przód
Aktorzy:	Użytkownik
Warunki początkowe:	Załadowane ścieżki
Opis realizowanej funkcji:	Szybsze osiąganie oczekiwanego momentu w nagraniu niż w trakcie odtwarzania
Sytuacje wyjątkowe:	Osiągnięty koniec ścieżek – zatrzymanie przewijania (i odtwarzania) i powrót do początku nagrania
Warunki końcowe:	Przewijane nagrania w przód
Powiązane	F7

wymagania:	

Identyfikator:	PU8	
Nazwa:	Przewijanie w tył	
Aktorzy:	Użytkownik	
Warunki początkowe:	Załadowane ścieżki	
Opis realizowanej funkcji:	Osiąganie wcześniejszego momentu w nagraniu niż aktualny	
Sytuacje wyjątkowe:	Osiągnięty początek ścieżek – zatrzymanie przewijania (i odtwarzania)	
Warunki końcowe:	Przewijanie nagrania w tył	
Powiązane wymagania:	F8	

Identyfikator:	PU9	
Nazwa:	Ustawienie ścieżki w tryb SOLO	
Aktorzy:	Użytkownik	
Warunki początkowe:	Wybrana ścieżka	
Opis realizowanej funkcji:	Ustawienie ścieżki w tryb SOLO – odtwarzanie tylko ścieżek ustawionych w tryb SOLO	
Sytuacje wyjątkowe:	Brak	
Warunki końcowe:	Ścieżka ustawiona w tryb SOLO	
Powiązane wymagania:	F9	

Identyfikator:	PU10	
Nazwa:	Wyciszenie ścieżki	
Aktorzy:	Użytkownik	
Warunki początkowe:	Wybrana ścieżka	

Opis realizowanej funkcji:	Wyciszenie ścieżki – w trakcie odtwarzania sumy ścieżek, ścieżka jest pomijana
Sytuacje wyjątkowe:	Brak
Warunki końcowe:	Ścieżka ustawiona w tryb wyciszenia
Powiązane wymagania:	F10

Identyfikator:	PU11	
Nazwa:	Dezaktywacja wszystkich trybów SOLO	
Aktorzy:	Użytkownik	
Warunki początkowe:	Załadowane ścieżki, co najmniej jedna ustawiona w tryb SOLO	
Opis realizowanej funkcji:	Anulowanie trybów SOLO na wszystkich ścieżkach	
Sytuacje wyjątkowe:	Brak	
Warunki końcowe:	Żadna ścieżka nie jest w trybie SOLO	
Powiązane wymagania:	F11	

Identyfikator:	PU12	
Nazwa:	Dezaktywacja wszystkich trybów wyciszenia	
Aktorzy:	Użytkownik	
Warunki początkowe:	Załadowane ścieżki, co najmniej jedna w trybie wyciszenia	
Opis realizowanej funkcji:	Anulowanie trybów wyciszenia na wszystkich ścieżkach	
Sytuacje wyjątkowe:	Brak	
Warunki końcowe:	Żadna ścieżka nie jest w trybie wyciszenia	
Powiązane wymagania:	F12	

Identyfikator:	PU13	
Nazwa:	Zmiana poziomu	
Aktorzy:	Użytkownik	
Warunki początkowe:	Wybrana ścieżka	
Opis realizowanej funkcji:	Płynna zmiana poziomu odtwarzania ścieżki	
Sytuacje wyjątkowe:	Brak	
Warunki końcowe:	Zmieniony poziom odtwarzania ścieżki	
Powiązane wymagania:	F13	

Identyfikator:	PU14	
Nazwa:	Ustawienie panoramy	
Aktorzy:	Użytkownik	
Warunki początkowe:	Wybrana ścieżka	
Opis realizowanej funkcji:	Ustalenie lokalizacji dźwięku dla wybranej ścieżki w przestrzeni wielokanałowej	
Sytuacje wyjątkowe:	Brak	
Warunki końcowe:	Ustalona lokalizacja dźwięku w przestrzeni wielokanałowej dla wybranej ścieżki	
Powiązane wymagania:	F14	

Identyfikator:	PU15	
Nazwa:	Dodanie pogłosu	
Aktorzy:	Użytkownik	
Warunki początkowe:	Wybrana ścieżka	
Opis realizowanej	Dodanie pogłosu do wybranej ścieżki	

funkcji:	
Sytuacje wyjątkowe:	Brak
Warunki końcowe:	Dodany pogłos do wybranej ścieżki
Powiązane wymagania:	F15

Identyfikator:	PU19	
Nazwa:	Ustawienie czasu pogłosu	
Aktorzy:	Użytkownik	
Warunki początkowe:	Wybrana funkcja dodawania pogłosu dla wybranej ścieżki	
Opis realizowanej funkcji:	Ustawienie czasu pogłosu dla wybranej ścieżki	
Sytuacje wyjątkowe:	Brak	
Warunki końcowe:	Ustawiony czas pogłosu	
Powiązane wymagania:	F19	

Identyfikator:	PU20	
Nazwa:	Ustawienie poziomu pogłosu	
Aktorzy:	Użytkownik	
Warunki początkowe:	Wybrana funkcja dodawania pogłosu do wybranej ścieżki	
Opis realizowanej funkcji:	Ustawienie stosunku dźwięku pogłosowego do bezpośredniego	
Sytuacje wyjątkowe:	Brak	
Warunki końcowe:	Ustawiony poziom pogłosu	
Powiązane wymagania:	F20	

Identyfikator:	PU16	
Nazwa:	Kompresja dynamiki	
Aktorzy:	Użytkownik	
Warunki początkowe:	Wybrana ścieżka	
Opis realizowanej funkcji:	Skompresowanie dynamiki sygnału wybranej ścieżki	
Sytuacje wyjątkowe:	Brak	
Warunki końcowe:	Ustawiony poziom kompresji dynamiki dla wybranej ścieżki	
Powiązane wymagania:	F16	

Identyfikator:	PU21	
Nazwa:	Ustawienie progu kompresji	
Aktorzy:	Użytkownik	
Warunki początkowe:	Wybrana funkcja kompresji dynamiki sygnału wybranej ścieżki	
Opis realizowanej funkcji:	Ustawienie progu kompresji sygnału wybranej ścieżki	
Sytuacje wyjątkowe:	Brak	
Warunki końcowe:	Ustawiony próg kompresji	
Powiązane wymagania:	F21	

Identyfikator:	PU22	
Nazwa:	Ustawienie stopnia kompresji	
Aktorzy:	Użytkownik	
Warunki początkowe:	Wybrana funkcja kompresji dynamiki sygnału wybranej ścieżki	
Opis realizowanej funkcji:	Ustawienie stopnia kompresji sygnału wybranej ścieżki	

Sytuacje wyjątkowe:	Brak
Warunki końcowe:	Ustawiony stopień kompresji
Powiązane wymagania:	F22

Identyfikator:	PU17	
Nazwa:	Korekcja częstotliwościowa	
Aktorzy:	Użytkownik	
Warunki początkowe:	Wybrana ścieżka	
Opis realizowanej funkcji:	Ustawienie krzywej korekcji częstotliwościowej sygnału wybranej ścieżki	
Sytuacje wyjątkowe:	Brak	
Warunki końcowe:	Ustawiona krzywa korekcji częstotliwościowej sygnału wybranej ścieżki	
Powiązane wymagania:	F17	

Identyfikator:	PU18	
Nazwa:	Eksport sumy do pliku	
Aktorzy:	Użytkownik	
Warunki początkowe:	Załadowane ścieżki	
Opis realizowanej funkcji:	Wyeksportowanie wszystkich ścieżek z uwzględnieniem wszystkich parametrów wpływających na odtwarzanie do jednego pliku o formacie dźwięku stereofonicznego lub 5 plików monofonicznych odpowiadających kanałom systemu 5.1	
Sytuacje wyjątkowe:	Brak	
Warunki końcowe:	Suma ścieżek wyeksportowana do pliku/plików	
Powiązane wymagania:	F18	

Identyfikator:	PU19	
Nazwa:	Otwarcie podręcznika użytkownika	
Aktorzy:	Użytkownik	
Warunki początkowe:	Uruchomiona aplikacja	
Opis realizowanej funkcji:	Otwarcie pliku .pdf zawierającego podręcznik użytkownika.	
Sytuacje wyjątkowe:	Brak	
Warunki końcowe:	Otwarty podręcznik użytkownika	
Powiązane wymagania:	F19	

Konceptualne diagramy klas

Konceptualne diagramy klas utworzone zostały na podstawie określenia rzeczowników modelowanej dziedziny.

Rzeczowniki dziedziny:

Związane z przetwarzaniem dźwięku:

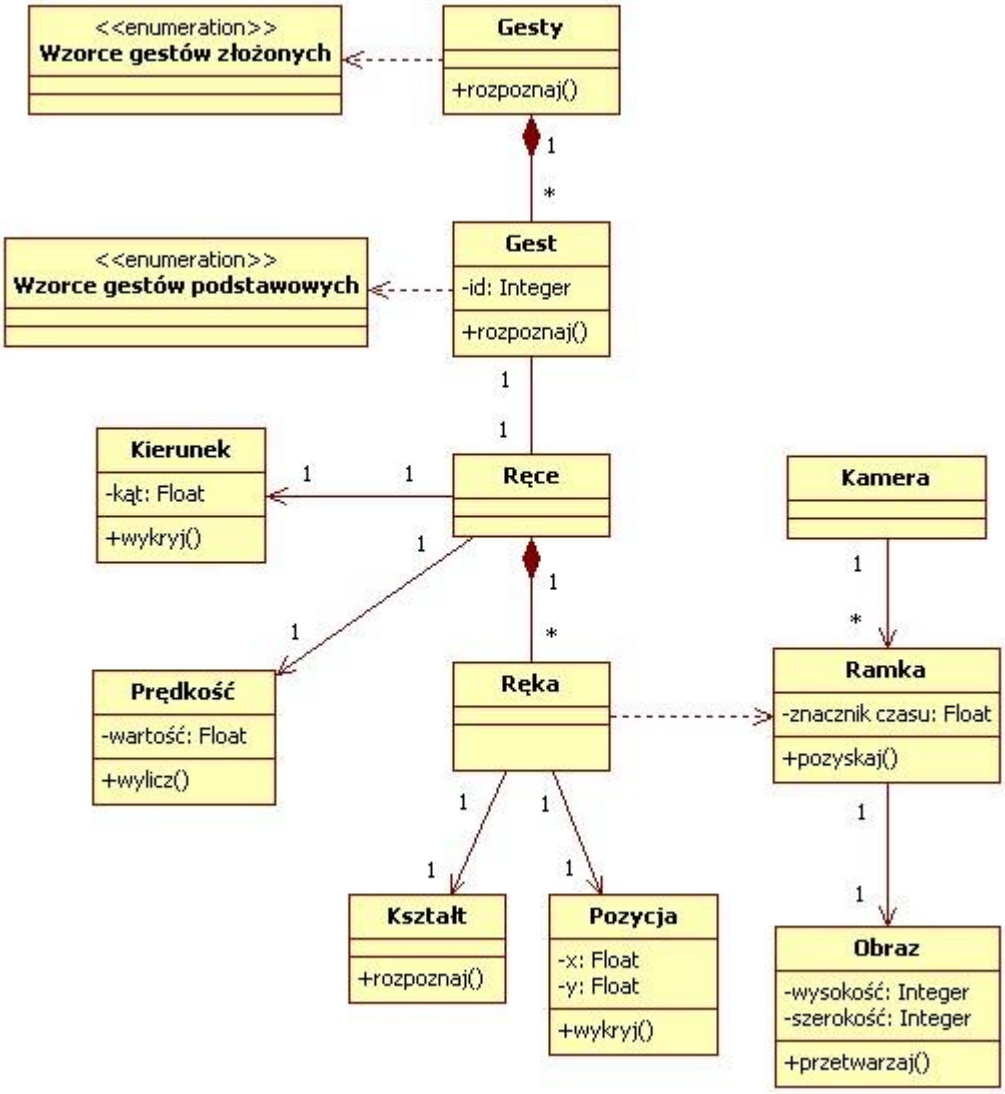
kompresor dynamiki, częstotliwość, pogłos, panorama, poziom dźwięku, ścieżka, suma, korektor częstotliwości, odtwarzacz, wysyłka, insert

Związane ze sterowaniem za pomocą gestów:

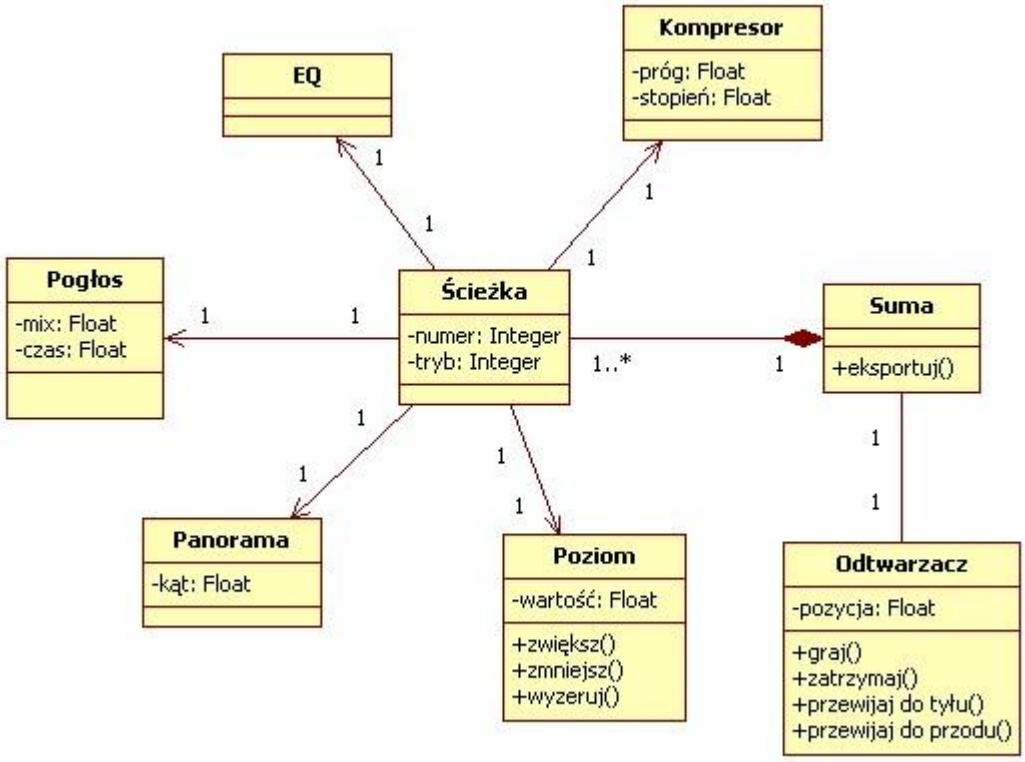
gest dynamiczny, gest statyczny, ręka, palec, dłoń, prędkość, kierunek, przyśpieszenie

Związane z przetwarzaniem obrazu:

kamera, ramka, obraz



Rys. G.2 Konceptualny diagram klas interfejsu umożliwiającego sterowanie aplikacją komputerową za pomocą gestów



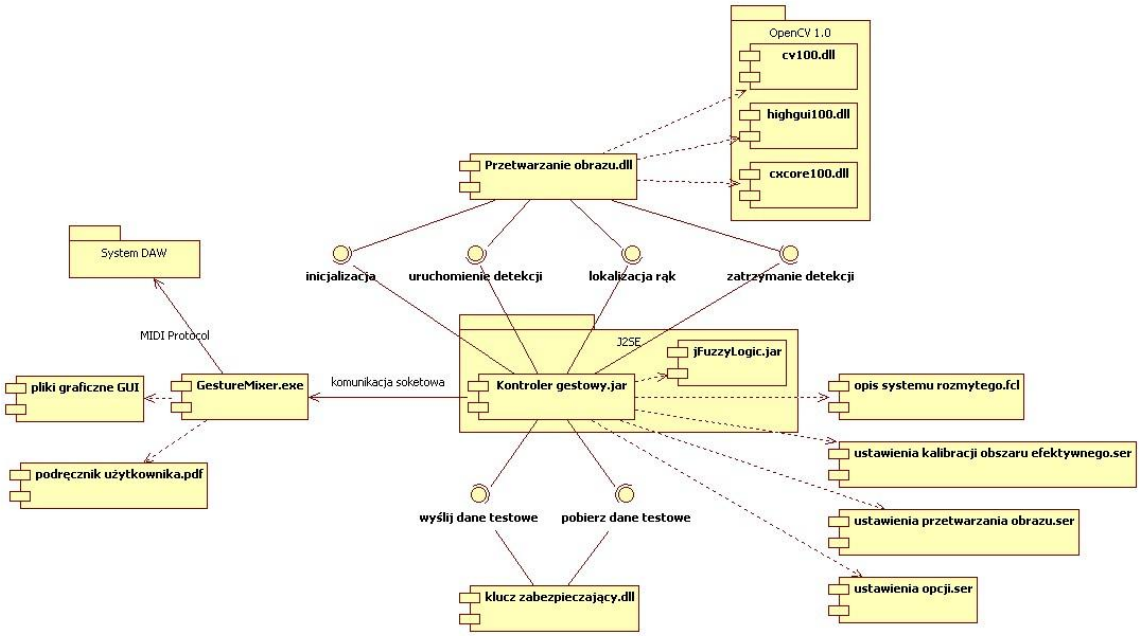
Rys. G.3 Konceptualny diagram klas aplikacji miksowania nagrań muzycznych

Architektura systemu

System wykorzystuje wcześniej opracowany interfejs sterowania komputerem za pomocą gestów. Część oprogramowania realizującą przetwarzanie obrazu i detekcję rąk zaimplementowano w środowisku Visual Studio 2008 C++ korzystając z biblioteki OpenCV w wersji 1.0. Nadrzędne metody udostępniono w postaci interfejsu JNI (ang. *Java Native Interface*) i skompilowano do biblioteki dll. Część odpowiedzialną za interpretację gestów i wykonywanie przypisanych gestom akcji systemowych oraz udostępniającą GUI zaimplementowano w środowisku Java SE. Wykorzystano pakiet jFuzzyLogic do konstrukcji systemu wnioskowania rozmytego służącego rozpoznawaniu gestów. Opis systemu, w języku FCL (ang. *fuzzy control language*), zawarto w pliku tekstowym. Dokonane za pośrednictwem GUI ustawienia przetwarzania obrazu, kalibracyjne oraz opcji, przechowywane są w plikach serializowanych klas. Interfejs zabezpieczono kluczem sprzętowym przed próbą nielegalnego powielania. Aplikacja komunikuje się z kluczem za pośrednictwem biblioteki dll.

Aplikacja miksowania dźwięku zostanie stworzona w środowisku Visual Studio 2008 w języku C++. Funkcjom realizującym przetwarzanie dźwięku przypisane zostaną skróty klawiaturowe. Wciśnięcie klawisza na klawiaturze bądź kombinacji klawiszy emulowane będzie przez kontroler gestowy, jako zdarzenie na wykonany gest lub gesty. Informacje o stanie obiektów w aplikacji miksowania dźwięku, który mógłby wpłynąć na sposób interpretowania gestów, przesyłane będą do kontrolera gestowego za pośrednictwem gniazd (komunikacja soketowa).

System powinien bezbłędnie realizować wyspecyfikowane wymagania działając w systemach operacyjnych: Microsoft Windows Vista, Windows 7.



Rys. G.4 Diagram komponentów systemu miksowania dźwięku za pomocą gestów

Dodatek H. Ankieta dla realizatorów

Badanie ankietowe interfejsu miksowania dźwięku za pomocą gestów

Staż pracy jako inżynier miksu: ……

System DAW używany na co dzień:
☐ Avid Pro Tools ☐ Steinberg Cubase / Nuendo
☐ Apple Logic ☐ Magix Samplitude
☐ Inny ……

1. Jak oceniasz intuicyjność gestów? …… (*1 – bardzo niska, 5 - bardzo wysoka)*
2. W przypadku wybrania odpowiedzi innej niż „5" w pyt. 1, napisz jakie zmiany wprowadził(a)byś w słowniku gestów, aby obsługa była bardziej intuicyjna?

   ……

   ……

   ……
3. Jak oceniasz wygodę użytkowania interfejsu w przypadku obsługi za pomocą gestów? …… (*1 – bardzo niska, 5 - bardzo wysoka)*
4. Jak oceniasz dokładność edycji wartości parametrów przy obsłudze za pomocą gestów? ……

   (*1 – zdecydowanie niewystarczająca, 5 – odpowiadająca systemowi DAW użytemu w trakcie testów*)
5. W przypadku oceny innej niż „5" w pyt. 4, edycję których parametrów uważasz za niewystarczająco dokładną?

Parametr	Ocena (1 - 5)
……	……
……	……
……	……
……	……
……	……
……	……
……	……

6. Czy w przypadku któregoś z systemów i sposobów obsługi zmysł wzroku był angażowany w mniejszym stopniu, tj. w większym stopniu można było skoncentrować się na dźwięku? ☐ TAK ☐ NIE

   (*W przypadku zaznaczenia odp. NIE przejdź do pytania nr 8)*
7. Mniejsze zaangażowanie zmysłu wzroku wystąpiło w przypadku:
   - ☐ systemu GrabSound obsługiwanego za pomocą gestów w trybie ograniczonego interfejsu graficznego

- ☐ systemu GrabSound obsługiwanego za pomocą gestów w trybie pełnego interfejsu graficznego
- ☐ systemu GrabSound obsługiwanego za pomocą myszy i klawiatury w trybie ograniczonego interfejsu graficznego
- ☐ systemu GrabSound obsługiwanego za pomocą myszy i klawiatury w trybie pełnego interfejsu graficznego
- ☐ systemu DAW

8. Czy byłeś/aś w stanie wykonać identyczne zgranie w obu systemach i we wszystkich wariantach? ☐ TAK ☐ NIE
9. Czy podczas miksowania w oparciu o system DAW zdarza Ci się wybierać wartości parametrów, które nie wywołują słyszalnej zmiany?
   ☐ TAK ☐ NIE
10. Przy zastosowaniu którego systemu/wariantu otrzymałeś/aś zgranie różniące się od pozostałych?
    - ☐ GrabSound / gesty / pełne GUI
    - ☐ GrabSound / gesty / ograniczone GUI
    - ☐ GrabSound / mysz i klawiatura / pełne GUI
    - ☐ GrabSound / mysz i klawiatura / ograniczone GUI
    - ☐ DAW
11. Co, Twoim zdaniem, było przyczyną otrzymania zgrania różniącego się od pozostałych?
    - ☐ większa dokładność
    - ☐ niewystarczająca dokładność
    - ☐ obecność informacji wizualnej odzwierciedlającej zmiany parametrów
    - ☐ brak informacji wizualnej odzwierciedlającej zmiany wartości parametrów
    - ☐ wygoda obsługi
    - ☐ brak wygody obsługi
    - ☐ możliwość jednoczesnej edycji dwóch parametrów
    - ☐ brak możliwości jednoczesnej edycji dwóch parametrów
    - ☐ inne ……
    - ☐ trudno powiedzieć
12. Uszereguj uzyskane zgrania w kolejności od najlepszego (1) do najgorszego (5):
    - ☐ GrabSound / gesty / pełne GUI
    - ☐ GrabSound / gesty / ograniczone GUI
    - ☐ GrabSound / mysz i klawiatura / pełne GUI
    - ☐ GrabSound / mysz i klawiatura / ograniczone GUI
    - ☐ DAW

13. Dodatkowe uwagi:

……

……

……

Dodatek I. Dodatek DVD

Zawartość płyty:

- Eksperymenty do rozdziału 7
  - wyniki miksowania
    - realizator 1
    - realizator 2
    - realizator 3
    - realizator 4
    - realizator 5
    - realizator 6
    - realizator 7
    - realizator 8
    - realizator 9
    - realizator 10
  - wyniki badania ankietowego