XX Międzynarodowe Sympozjum Nowości w Technice Audio i Wideo NTAV2024

Wrocław, 17-19 października 2024

Katedra Akustyki, Multimediów i Przetwarzania Sygnałów

PAN POLSKA AKADEMIA NAUK

Polish Audio Engineering Society

Materiały konferencyjne

ORGANIZATORZY ▪ ORGANIZERS

Katedra Akustyki, Multimediów i Przetwarzania Sygnałów
Polskie Towarzystwo Akustyczne oddział we Wrocławiu
Polska Sekcja Audio Engineering Society
Polska Akademia Nauk
Politechnika Wrocławska

KOMITET NAUKOWY ▪ SCIENTIFIC COMMITTEE

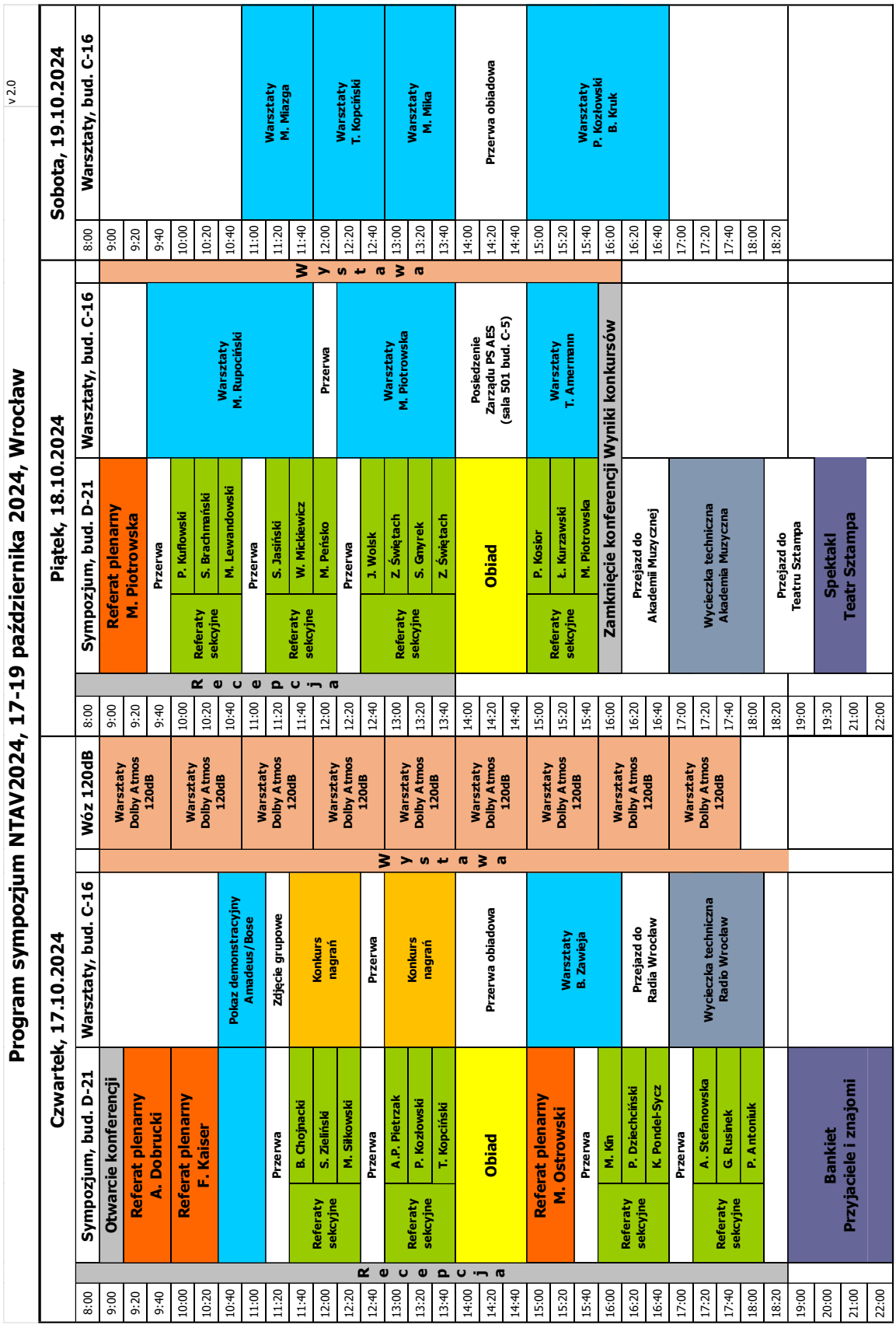
Przewodniczący: Krzysztof Opieliński
Andrzej Brzoska
Andrzej Czyżewski
Andrzej Dobrucki
Tadeusz Kamisiński
Piotr Kleczkowski
Bożena Kostek
Mirosław Meissner
Andrzej Miśkiewicz
Janusz Piechowicz
Adam Pilch
Anna Preis
Tomira Rogala
Ewa Skrodzka
Bogusław Szlachetko
Andrzej Wicher
Jerzy Wiciak
Sławomir Zieliński
Jan Żera

KOMITET ORGANIZACYJNY ▪ ORGANIZING COMMITTEE

Przewodniczący:	Przemysław Plaskota *(korekty)*
Sekretarz:	Michał Łuczyński *(redakcja)*
Skarbnik:	Paweł Dziechciński
Członkowie:	Bartłomiej Kruk
	Maurycy J. Kin
	Piotr Kozłowski
	Andrzej Lewandowski
	Agnieszka Wielgus
	Romuald Bolejko
	Zbigniew Świetach
	Agnieszka Paula Pietrzak

WSPIERAJĄ NAS

softserve


Program sympozjum NTAV2024, 17-19 października 2024, Wrocław
v 2.0
Czwartek, 17.10.2024
Sympozjum, bud. D-21
Warsztaty, bud. C-16
Wóz 120dB
Otwarcie konferencji
Referat plenarny A. Dobrucki
Referat plenarny F. Kaiser
Pokaz demonstracyjny Amadeus/Bose
Przerwa
Zdjęcie grupowe
Referaty sekcyjne
B. Chojnacki
S. Zieliński
M. Sikowski
Konkurs nagrań
Przerwa
A.P. Pietrzak
P. Kozłowski
T. Kopciński
Konkurs nagrań
Obiad
Przerwa obiadowa
Referat plenarny M. Ostrowski
Przerwa
M. Kin
P. Dziechciński
K. Pondel-Sycz
Warsztaty B. Zawieja
Przejazd do Radia Wrocław
Przerwa
A. Stefanowska
G. Rusinek
P. Antoniuk
Wycieczka techniczna Radio Wrocław
Bankiet Przyjaciele i znajomi
Warsztaty Dolby Atmos 120dB
Piątek, 18.10.2024
Sympozjum, bud. D-21
Warsztaty, bud. C-16
Referat plenarny M. Piotrowska
Przerwa
Referaty sekcyjne
P. Kuflowski
S. Brachmański
M. Lewandowski
Warsztaty M. Rupociński
Przerwa
Referaty sekcyjne
S. Jasiński
W. Mickiewicz
M. Pełsko
Przerwa
Przerwa
Referaty sekcyjne
J. Wolsk
Z. Świętach
S. Gmyrek
Z. Świętach
Warsztaty M. Piotrowska
Obiad
Posiedzenie Zarządu PS AES (sala 501 bud. C-5)
Referaty sekcyjne
P. Kosior
Ł. Kurzawski
M. Piotrowska
Warsztaty T. Amermann
Zamknięcie konferencji Wyniki konkursów
Przejazd do Akademii Muzycznej
Wycieczka techniczna Akademia Muzyczna
Przejazd do Teatru Sztampa
Spektakl Teatr Sztampa
Sobota, 19.10.2024
Warsztaty, bud. C-16
Warsztaty M. Miazga
Warsztaty T. Kopciński
Warsztaty M. Mika
Przerwa obiadowa
Warsztaty P. Kozłowski B. Kruk

SPIS TREŚCI

Paper ID: **1**

SYNTEZA DŹWIĘKU PRZESTRZENNEGO Z WYKORZYSTANIEM ZMODYFIKOWANEJ BAZY HRTF POMIARÓW WYKONANYCH NA POLITECHNICE WROCŁAWSKIEJ

Zbigniew Świętach, Przemysław Plaskota
Katedra Akustyki, Mutlimediów i Przetwarzania Sygnałów, Wydział Elektroniki, Fotoniki i Mikrosystemów, Politechnika Wrocławska
Autor korespondencyjny: **Zbigniew Świętach**, *zbigniew.swietach@pwr.edu.pl*

Słowa kluczowe: lokalizacja źródeł dźwięku, HRTF, sysnteza pola akustycznego

WPROWADZENIE

W niniejszej pracy przedstawiono wyniki syntezy dźwięku przestrzennego na podstawie zadanego monofonicznego sygnału dźwiękowego. W tym celu wykorzystano zbiór przestrzennie skorelowanych odpowiedzi impulsowych lub równoważnie transmitancji przestrzennych mierzonych dookólnie względem środka głowy potencjalnego słuchacza (HRTF – Head Related Transfer Functions) [1, 2]. We wzmiankowanej metodzie, fizyczne nieograniczone czasowo odpowiedzi impulsowe przybliża się odpowiednio dobranymi skończonymi ciągami próbek (dyskretyzacja odpowiedzi impulsowych).

Do celów syntezy dźwięku przestrzennego wykorzystano środowisko obliczeniowe Matlab. Wybór Matlaba podyktowany jest prostotą kodu źródłowego, który jest plikiem tekstowym. Ponadto sposób zapisu problemu technicznego czy matematycznego w Matlabie jest intuicyjny i niewiele odbiega od zapisu takiego problemu przy użyciu standardowej notacji matematycznej.

Niniejsza praca jest kontynuacją prac nad bazą pomiarów HRTF wykonanych na Politechnice Wrocławskiej. Oryginalna baza zawierała błędy związane z akwizycją sygnałów pomiarowych. W tej pracy omawiany jest sposób modyfikacji oryginalnej bazy HRTF, który umożliwił eliminację większości błędów.

MODYFIKACJA BAZY HRTF I SYNTEZA DŹWIĘKU PRZESTRZENNEGO

Przestrzenne odpowiedzi impulsowe HRTF zostały uprzednio zmierzone i były dostępne w formacie xml [4]. Ze względu na dużą ilość danych niezbędnych do przeprowadzenia syntezy dźwięku, wybrano metodę modyfikowania pomiarów powiązanych z pojedynczym słuchaczem. Po wyznaczeniu widm DFT dla kilkunastu kierunków, stwierdzono, że widma nie zostały prawidłowo ograniczone przed procesem próbkowania, tzn. nie zastosowano odpowiednich analogowych filtrów antyaliasingowych. Zapisane próbki sygnałów są obarczone pewnymi błędami aliasingu i tego nie można wyeliminować za pomocą metod post-processingu. Można tylko mieć nadzieję, że efekt końcowy, czyli dźwięk przestrzenny będzie jednak dobrze odbierany przez słuchacza.

Pozostałe mankamenty oryginalnych HRTF tzn. przydźwięki występujące przy około 50 Hz oraz 100 Hz zostały odfiltrowane, a pasmo oryginalnych HRTF zostało ograniczone do około [0.2, 8.0] kHz. Dla HRTF mierzonych w zakresie azymutu [15, 180] stopni wystąpił błąd przesunięcia ciągów próbek

dla lewego i prawego kanału. Błąd ten wyeliminowano metodą „porównawczą". Porównywano opóźnienia w obydwu kanałach przy poprawnie przeprowadzonej akwizycji sygnału, np. dla azymutu - 15 stopni i następnie przesuwano źle opóźnione ciągi próbek względem siebie aby uzyskać takie same opóźnienia dla azymutu +15 stopni. W większości przypadków należało przesunąć źle opóźniony ciąg próbek o połowę liczby próbek. Tak zmodyfikowana baza HRTF została użyta do syntezy dźwięku przestrzennego.

Rozważano syntezę dźwięku przestrzennego w przypadkach, gdy źródło dźwięku przemieszcza się wirtualnie, jak i gdy źródło dźwięku znajduje się w określonym położeniu względem słuchacza. Jeżeli źródło dźwięku przemieszcza się wirtualnie względem słuchacza, wówczas w obydwu metodach istotna jest długość ramki czasowej pomiędzy dwoma kolejnymi zmianami położenia źródła dźwięku. Wirtualny ruch źródła dźwięku realizowany jest przez splatanie sygnału monofonicznego źródła dźwięku z kolejno wybraną odpowiedzią impulsową ze zbioru dostępnych HRFT.

W pracy [3] rozważa się możliwość predykcji długości wzmiankowanej ramki czasowej, jednak aktualnie w niniejszej pracy nie wykorzystuje się tej metody. Teraz ustala się arbitralnie długość ramki czasowej dla całego procesu syntezy dźwięku przestrzennego. Podczas konferencji dostępne będą przykłady omawianej syntezy dźwięku przestrzennego w formie plików *.wav. Podsumowując, wydaje się, że modyfikacja bazy HRTF PWr powiodła się i wzmiankowana baza pomiarów może być wykorzystywana do syntezy dźwięku przestrzennego.

LITERATURA

[1] DOBRUCKI A., PLASKOTA P., PRUCHNICKI P., PEC M., BUJACZ M., STRUMILLO P. *Measurement System for Personalized Head-Related Transfer Functions and Its Verificationby Virtual Source Localization Trials with Visually Impaired and Sighted Individuals*. J. Audio Eng. Soc., Vol. 58, No. 9, 2010 September.

[2] CHENG C.I., WAKEFIELD G.H. *Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space*. J Audio Eng Soc, Vol 49, No 4, 2001 April.

[3] PAUSCH F., DOMA S., FELS J. *Hybrid multi-harmonic model for the prediction of interaural time differences in individual behind-the-ear hearing-aid-related transfer functions*. Acta Acustica 2022, 6, 34.

[4] PLASKOTA P., STASIAK J. *Baza danych zawierająca wyniki pomiarów hrtf w formacie xml*. Postępy Akustyki, Gliwice 2017.

Paper ID: **2**

OCENA JAKOŚCI DŹWIĘKU MOWY SYNTETYCZNEJ

Stefan Brachmański, Maurycy Kin, Piotr Kozłowski
Katedra Akustyki, Mutlimediów i Przetwarzania Sygnałów, Wydział Elektroniki, Fotoniki i Mikrosystemów, Politechnika Wrocławska
Autor korespondencyjny: **Maurycy Kin**, *maurycy.kin@pwr.edu.pl*

Słowa kluczowe: jakość sygnału mowy, synteza mowy, wyrazistość logatomowa

1. WPROWADZENIE

Dla zapewnienia dobrej komunikacji między osobami niezbędnych jest szereg czynników warunkujących jakość dźwięku mowy. Oprócz typowych parametrów odzwierciedlających zrozumiałość i wyrazistość mowy, niezbędnym wydaje się być zapewnienie odpowiedniej naturalności wypowiedzi [1]. Bardzo ważnym elementem mowy syntetycznej jest więc jej ogólna jakość oraz zrozumiałość. Zrozumiałość mowy jest jednym z podstawowych parametrów jakościowych transmisji sygnału mowy zarówno w analogowych, jak i cyfrowych sieciach telekomunikacyjnych, a także w salach audytoryjnych i stosowanych w nich systemach dźwiękowych, elektroakustycznych systemach ostrzegawczych, syntezatorach oraz w aparatach słuchowych. Zrozumiałość mowy nie jest tym samym parametrem, co wierność i naturalność mowy. Na przykład sygnał mowy emitowany w systemach ostrzegawczych może nie brzmieć przyjemnie, jednak komunikaty ostrzegawcze muszą być przekazywane w skuteczny, zrozumiały sposób.

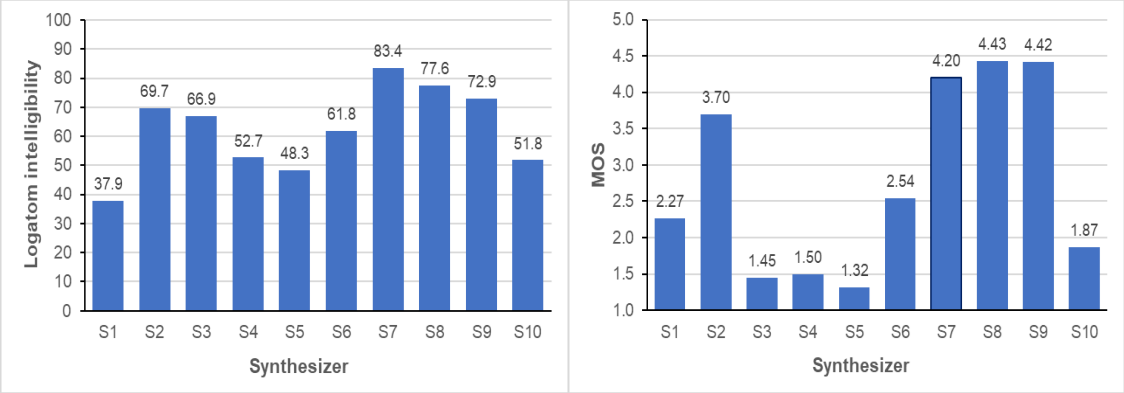
Ocena jakości i zrozumiałości mowy tak naturalnej, jak i syntetycznej może być dokonywana za pomocą różnego rodzaju metod subiektywnych oraz obiektywnych [2,3]. W niniejszej pracy postanowiono zweryfikować przydatność metody oceny jakości mowy syntetycznej za pomocą kryterium jakościowego skalowania absolutnego (Absolute Category Rating) oraz metody zawierającej aspekty zrozumiałościowe, tzn. ocenę wyrazistości logatomowej.

2. EKSPERYMENT

W badaniu wzięło udział 30 osób z dobrym słuchem, w wieku 20-25 lat, doświadczonych w badaniach nad oceną jakości dźwięku. Do badań wykorzystano listy logatomowe zawierające 100 logatomów (test wyrazistości) oraz listy zdaniowe (ocena jakości). Do wygenerowania dźwięków mowy wykorzystano dziesięć syntezatorów:

Syntezator WP (głos męski), 2. Realspeak (głos żeński - Agata), 3. Syntalk (głos męski), 4. eSpeak (głos męski), 5. eSpeak (głos żeński), 6. mySimpleSynth (głos żeński), 7. Acapela (głos żeński - Ania), 8. Expressivo (głos męski - Jacek), 9. Expressivo (głos żeński - Ewa) oraz 10. Dant Free (głos męski).

Sygnał akustyczny rejestrowano jednokanałowo w formacie PCM z szybkością próbkowania wynoszącą 16000 próbek/s i rozdzielczością 16-bitów. Badania przeprowadzono metodą odsłuchu słuchawkowego. Uzyskane wyniki poddano weryfikacji statystycznej w zakresie homogeniczności wariancji oraz zgodności rozkładów, a następnie uśredniono w grupie słuchaczy. Na Rys. 1 przedstawiono w formie procentowej wyniki wyrazistości logatomowej oraz oceny jakości MOS badanych syntezatorów.



Rys. 1. Wyniki wyrazistości logatomowej oraz ogólnej oceny jakości badanych syntezatorów.

Otrzymane wartości procentowe wyrazistości logatomowej zawierają się w przedziale 37,9 – 83,4 %. Można więc stwierdzić, że pod względem wyrazistości badane urządzenia zostały ocenione jako dobre i bardzo dobre, zgodnie z przyjętymi kryteriami [4]. Najwyższe oceny jakości otrzymały syntezatory S7, S8 oraz S9 – we wszystkich tych przypadkach można mówić o jakości lepszej niż dobra (MOS > 4,0). Także jakość sygnału mowy generowanej przez syntezator S2 (Realspeak generujący głos żeński) została uznana jako nieco tylko gorsza od dobrej (MOS = 3,7). Natomiast jakość dźwięku mowy generowanej przez pozostałe syntezatory została oceniona znacznie gorzej – (MOS < 2,54), co oznacza, że mogą być one stosowane w specjalnych aplikacjach, gdzie jakość dźwięku nie jest kluczowa.

Uzyskane wyniki wskazują, że wysoka wyrazistość logatomowa nie gwarantuje wysokiej jakości dźwięku, a spowodowane jest to brakiem cech prozodycznych dźwięków mowy, nie uwzględnieniem intonacji, a tym samym naturalnych emocji zawartych w głosie ludzkim.

LITERATURA

[1] BRACHMAŃSKI S., Wybrane zagadnienia oceny jakości transmisji sygnału mowy, Wyd. Politechniki Wrocławskiej, 2015.

[2] KITAWAKI N., NAGABUCHI H., *"Quality assessment of speech coding and speech synthesis systems"*, IEEE Communications Magazine, October, 36 – 44, 1988.

[3] KIN M., BRACHMAŃSKI S., *"Quality assessment of musical and speech signals broadcasted via Single Frequency Network DAB+"*. Int. Journal of Electronics and Telecommunications, vol. 66, no. 1, 139 – 144, 2020.

[4] MYŚLECKI W., MAJEWSKI W., *"Relations between subjective and objective measures of speech transmission quality evaluation"*, Proc. of 6th FASE Symposium, Sopron, Budapest, 137-141, 2-6 September 1986.

Paper ID: **3**

OBIEKTYWNA OCENA ZROZUMIAŁOŚCI MOWY KODEKÓW STRATNYCH WYSOKIEJ JAKOŚCI METODĄ STIPA

Paweł Dziechciński
Katedra Akustyki, Mutlimediów i Przetwarzania Sygnałów, Wydział Elektroniki, Fotoniki i Mikrosystemów, Politechnika Wrocławska
Autor korespondencyjny: **Paweł Dziechciński**, *pawel.dziechcinski@pwr.edu.pl*

W pracy przedstawiono wyniki badań wpływu kodowania stratnego na wskaźnik transmisji mowy dotyczący systemów rozgłoszeniowych (STIPA). W analizach uwzględniono typ kodeka, jego dostawcę, szybkość transmisji strumienia danych na wyjściu kodera, częstotliwość próbkowania sygnału, liczbę kodowanych kanałów, poziomu sygnału cyfrowego oraz inne parametry specyficzne dla danego systemu kompresji. W sumie przeanalizowano ponad 8000 kombinacji tych parametrów. Na podstawie uzyskanych wyników wyznaczono graniczne wartości przepływności fonicznych kodeków stratnych nie zniekształcających sygnału STIPA, na potrzeby wykorzystania ich jako źródła sygnału pomiarowego.

Słowa kluczowe: zrozumiałość mowy; wskaźnik transmisji mowy; STIPA; kompresja stratna

Paper ID: **4**

OCENA JAKOŚCI VIDEO METODAMI SINGLE STIMULUS IMPAIRMENT SCALE I DOUBLE STIMULUS IMPAIRMENT SCALE

Stefan Brachmański [1], Janusz Klink [2], , Michał Łuczyński [1]
[1] Katedra Akustyki, Mutlimediów i Przetwarzania Sygnałów, Wydział Elektroniki, Fotoniki i Mikrosystemów, Politechnika Wrocławska,
[2] Katedra Telekomunikacji i Teleinformatyki, Wydział Informatyki i Telekomunikacji, Politechnika Wrocławska
Autor korespondencyjny: **Stefan Brachmański**, *stefan.brachmanski@pwr.edu.pl*

Słowa kluczowe: subiektywna ocena jakości wideo; metoda Single-Stimulus, metoda Double-Stimulus Impairment Scale (DSIS) method, kodowanie video, H264, H265

WPROWADZENIE

Transmisja video realizowana jest z wykorzystaniem różnych technik kodowania. International Telecommunication Union (ITU) zaleca korzystanie ze standardu H.264 [1] i nowszego H.265[2]. Wpływ na postrzeganie jakości video przez widza mają między innymi szybkość bitowa oraz rozdzielczość obrazu [3], [4], [5], [6].

Celem prezentowanej pracy było:

wykonanie oceny jakości video zalecanymi przez International Telecommunication Union metodami: Single-Stimulus (SS) i Double-Stimulus Impairment Scale (DSIS),

zbadanie wpływu na ocenę jakości sygnału video przez młodego użytkownika końcowego takich parametrów jak technika kodowania (H.264 i H.265), rozdzielczość (1280 x 720 i 1920 x 1080) i szybkość bitowa,

porównanie ocen jakości video otrzymanych z wykorzystaniem metody Single-Stimulus (SS) i metody Double-Stimulus Impairment Scale (DSIS).

2. EKSPERYMENT

Metoda Single-Stimulus (SS) [7] polega na ocenie przez obserwatorów jakości obrazu lub sekwencji obrazów (video) w pięciostopniowej skali. W tej metodzie prezentowane są testowe sekwencje wideo bez sekwencji odniesienia. Po każdej prezentowanej sekwencji testowej widz dokonuje oceny jakości obejrzanej sekwencji wideo, gdzie ocena 5 oznacza doskonałą jakość, a 1 - niedostateczną.

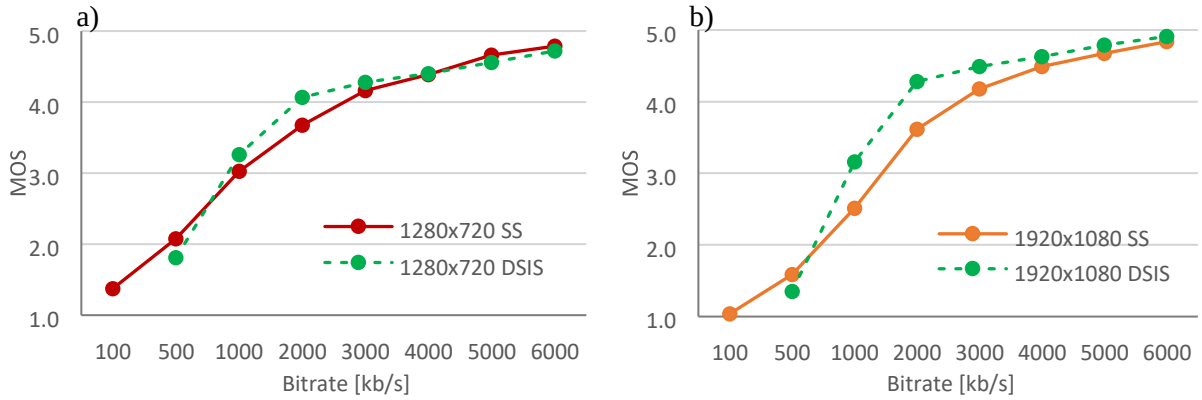
Metoda Double-Stimulus Impairment Scale (DSIS) [7] zalecana przez International Telecommunication Union jest alternatywnym rozwiązaniem dla metody SS. W metodzie DSIS, obserwatorowi prezentowane są dwie sekwencje wideo. Pierwszy bodziec jest sekwencją odniesienia, natomiast drugi jego zniekształconą wersją. Celem tej metody jest porównanie jakości zniekształconego obrazu video w odniesieniu do obrazu odniesienia. Obserwator podaje stopień pogorszenia jakości drugiego, zniekształconego obrazu, w pięciostopniowej Categorical Rating Scale (CRS), gdzie 5 oznacza niedostrzegalne zniekształcenia i zakłócenia, a 1 – bardzo dokuczliwe.

W eksperymencie udział wzięły dwie ekipy, jedna w pomiarach metodą SS, a druga – DSIS. W obu rodzajach pomiarów grupy obserwatorów tworzyli studenci Politechniki Wrocławskiej w wieku 20 – 21 lat o prawidłowej ostrości widzenia i poprawnym rozróżnianiu kolorów. W pomiarach metodą DSIS obserwatorów podzielono na -2 grupy. Liczebność poszczególnych grup była różna i tak dla kodowania

H.264 wynosiła 45 osób, a dla H.265 – 35 osób. Z kolei w pomiarach metodą SS grupa obserwatorów liczyła 80 osób zarówno dla techniki kodowania H.264 jak i H.265.

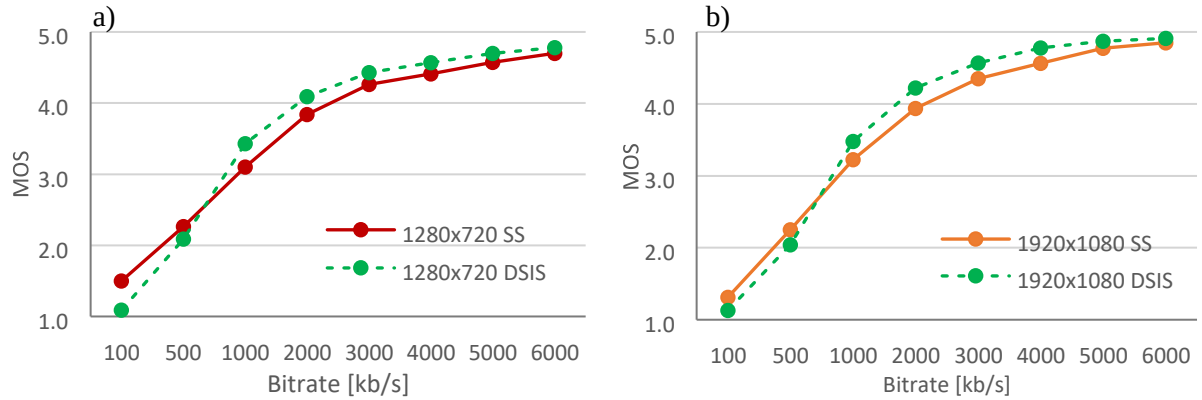
Testowym materiałem odniesienia była 20 sekundowa sekwencja video (bez dźwięku) o rozdzielczości 1920 x 1080 w formacie avi [3]. Sekwencja wideo zawierała sceny dynamiczne w postaci fragmentu startu wyścigów konnych. Sekwencję tą poddano kodowaniu H.264 i H.265 z różnymi szybkościami bitowymi i różną rozdzielczością.

Uśrednione wyniki oceny jakości wideo metodą SS i DSIS dla techniki kodowania H.264 przedstawiono na rysunku 1.



Rys. 1. Wyniki subiektywnej oceny jakości (DMOS i MOS) dla wideo kodowanego w standardzie H.264 jako funkcja szybkości transmisji bitów dla rozdzielczości 1280x720 (a) i 1920x1080 (b).

Uśrednione wyniki oceny jakości wideo metodą SS i DSIS dla techniki kodowania H.265 przedstawiono na rysunku 2.



Rys. 2. Wyniki subiektywnej oceny jakości (DMOS i MOS) dla wideo kodowanego w standardzie H.265 jako funkcja szybkości transmisji bitów dla rozdzielczości 1280x720 (a) i 1920x1080 (b).

Otrzymane wyniki pokazują na dużą zbieżność ocen otrzymanych obydwoma metodami dla kodowania H.265 dla analizowanych rozdzielczości i szybkości bitowych. W przypadku kodowania H.264 zgodność wyników dotyczy rozdzielczości 1280x720, natomiast dla rozdzielczości 1920x1080 zbieżność wyników obserwuje się dla szybkości bitowych powyżej 3000 kb/s.

LITERATURA

[1] ITU-T Rec. H.264, “Audiovisual and multimedia systems: Infrastructure of audiovisual services-Coding of moving video, Advanced video coding for generic audiovisual services,” 2021.
[2] ITU-T Rec. H.265, “High efficiency video coding ITU Publications International Telecommunication Union,” 2023.
[3] S. BRACHMAŃSKI, J. KLINK, “Subjective Assessment of the Quality of Video Sequences by the Young
[4] Viewers,” in 30 th International Conference on Software, Telecommunications and Computer Networks (Soft-COM 2022), Split: FESB, University of Split, pp. 1–6, 2022,
[5] J. KLINK, S. BRACHMAŃSKI, M. ŁUCZYŃSKI, „Assessment of the Quality of Video Sequences Performed by Viewers at Home and in the Laboratory”. Applied Sciences, vol. 13, nr 8, 5025, 2023, DOI: 10.3390/app13085025
[6] J. KLINK, S. BRACHMAŃSKI, M. ŁUCZYŃSKI, “Video Quality Modelling—Comparison of the Classical and Machine Learning Techniques”, Applied. Science, vol. 14, nr 16, 7029, 2024, https://doi.org/10.3390/app14167029
[7] F. M. MOSS, K. WANG, F. ZHANG, R. BADDELEY,D. R. BULL, “On the optimal presentation duration for subjective video quality assessment”. *IEEE Transactions on Circuits and Systems for Video Technology*, *26*(11), 1977-1987, 2015, DOI: 10.1109/TCSVT.2015.2461971
[8] ITU-R BT 500-14, “Methodologies for the subjective assessment of the quality of television images,” 2020.

Paper ID: **5**

COMPARISON OF AMBISONIC AND OBJECT-BASED SPATIAL SOUND RECORDING TECHNIQUES

Patryk Kosior, Bartłomiej Mróz
Katedra Systemów Multimedialnych, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska
Corresponding author: **Patryk Kosior**, *s180417@student.pg.edu.pl*

Keywords: higher-order ambisonics, object-based audio, MUSHRA, spatial audio, room divergence effect, immersion, localizability

1. INTRODUCTION

This article presents a comparison of spatial sound recording techniques based on scene-based and object-based ambisonic audio. The study aimed to assess the relationships between ambisonic orders of higher-order ambisonic scene and spot microphones encoded into ambisonics.

2. RECORDING STAGE

The audio material was obtained during an acoustic recording of instruments: piano, viola, electric bass guitar and a vocalist. Every musician was located on a semicircular plan, similarly as during a rehearsal or a recital. In the middle of the created plan a higher-order ambisonic microphone (namely, Zylia ZM-1) was placed. Additionally, all instruments were given spot microphones. The musicians performed an acoustic cover of a popular pop song.

3. POST-PRODUCTION STAGE

During the post-production phase, some audio signals processing operations was conducted. For example, the loudness of spot-recorded tracks was normalized so that the instruments recorded with spot microphones matched the loudness in the ambisonic-recorded scene. This normalization was done with the use of higher-order beams obtained from the ambisonic recording pointed to the locations of the respective instruments. Other important operation was phase matching. Some of the corresponding signals (e. g. piano signal from ambisonic microphone and piano signal form spot microphones) were not in phase. Therefore, a delay plugin was used to delay one of these signals.

Regarding the sound objects, they were created from signals recorded with spot microphones via simple ambisonics panning on a 3D sphere. Then, such obtained ambisonic signal was rendered with various spatial resolution: 1st-order ambisonics (1OA), 3rd-order (3OA) and 5th-order ambisonics (5OA). As for the scene-based signal, the ambisonic array is capable of only 3rd-order ambisonics; therefore, it was upscaled to the 5th-order. Thus, both object based and scene-based ambisonic signals consisted of 1st, 3rd and 5th order ambisonics.

Furthermore, reach of the scene-based variants was mixed with each of the object-based variants, resulting in 9 variants of the stimuli presented during the listening test. Additionally, three borderline cases were added, where no fusion was involved: 3rd-order ambisonic scene, 5th-order ambisonic (up-scaled) scene and 5th-order ambisonic-encoded objects. In total, 12 stimuli were presented during the listening experiment procedure.

4. LISTENING TEST STAGE

As a listening experiment, a MUSHRA-like test was conducted on a panel of expert listeners. The test procedure was conducted using open-back electrostatic reference headphones. No visual cues were provided. The binaural rendering was provided with three degrees of freedom via a headtracking device. The test was conducted twice in different acoustical environments: first room had acoustic treatment and short reverberation time (control room of a recording studio) and the second room with longer reverberation time (a shoebox-shaped classroom). Two different environments were necessary to scrutinize whether the room divergence effect occurred. Each test composed of two trials: in the first trial stimuli were rated in terms of immersion (which stimulus had the most immersive character) and in the second trial – in terms of localizability quality of each sound source. Each stimulus was rated on a scale 0-100. Participants could listen to each stimulus with any order and as many times as they wanted. In each trial the starting order of stimuli was randomized.

During the test, a repeated-measures design with fully crossed treatments was utilized. Following the test, statistical analysis was executed using the linear mixed models' approach with additional post-hoc statistical analysis to find statistically crucial differences between results.

5. RESULTS AND CONLUSIONS

Despite the acoustical difference, results from both rooms were similiar, indicating that change of acoustic conditions did not have a significant impact on ratings. Moreover, in both tests the best ratings were given to standalone 5th-order ambisonic-encoded objects variant. It is important to mention that the material was not mixed in any way (every stimulus was just a raw recording). As reported by some participants, since the stimuli presented an acoustic version of a pop song, it was expected for the vocals to be in the middle of stereo panorama and to have a dominant character. This might be one of the reasons for such high ratings for almost non-reverberant stimulus – namely the 5th -order ambisonic-encoded objects. Another aspect could be the fact that the spot microphones used during the recording were of very high quality. The combined quality of the spot microphones could outperform the tonal balance of the MEMS-based ambisonic microphone, which could be an important factor in the panel of audio engineering experts. In the subsequent investigation, using the same model of spot microphone for each application could be considered, ensuring their cumulative quality provides a similar level of quality as the ambisonic microphone.

Future work in this direction could include a more diverse set of music styles. Classical music should be especially considered due to high importance of acoustic reverberation in this particular style, which might be more suitable for immersion and envelopment ratings.

Additionally, a different approach for obtaining the stimuli could be considered. Instead of recording a live session, a set of loudspeakers could mimic the musicians. With the use of publicly available databases of stems and tracks, a much larger and a much more reliable set of stimuli could be constructed. Also, instead of headtracked-binaural playback, a loudspeaker based evaluation should improve the overall experimental procedure. However, a loudspeaker system with sufficient number of channels often is integrated with the room, so the room divergence effect might be difficult to mitigate.

Paper ID: **6**

SOUNDMAP - BAZA FUNKCJI PRZENOSZENIA SŁUCHAWEK (HPTF)

Szymon Jasiński, Bartłomiej Mróz, Bożena Kostek
Katedra Systemów Multimedialnych, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska,
Laboratorium Akustyki Fonicznej, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska
Autor korespondencyjny: **Szymon Jasiński**, *jasszy38@gmail.com*

Słowa kluczowe: funkcja przenoszenia słuchawek (HpTF), funkcja przenoszenia głowy (HRTF), dźwięk przestrzenny, dźwięk 3D, lokalizacja dźwięku, percepcja fal akustycznych, stanowisko pomiarowe, rejestracja, baza danych

W publikacji przedstawiono bazę funkcji przenoszenia słuchawek nausznych i wokół usznych zamkniętych (HpTF – ang. Headphone Transfer Function) dla wybranych czternastu modeli słuchawek i jednej pary słuchawek półotwartych. Zmierzono kilka modeli, które już znajdują się w innych znanych bazach [1,2,3,4] czy narzędziach do renderu binauralnego, jak np. Virtuoso [5]. Jednakże, typowe bazy HpTF opierają się na profesjonalnych, studyjnych modelach słuchawek; zaprezentowana w niniejszym artykule baza składa się w większości z modeli konsumenckich, których filtry HpTF nie zostały zmierzone wcześniej.

Zawarto również opis poszczególnych etapów realizacji bazy. W szczególności zdefiniowano założenia projektowe, zawarto opis wykorzystanego sprzętu i oprogramowania oraz środowiska pomiarowego. Wyszczególniono między innymi specyfikację użytych mikrofonów, symulatora głowy i torsu oraz uruchomionego skryptu. Opisano schemat podłączenia pomiarowego wraz z algorytmem przeprowadzania pomiarów.

Pomiary wykonano w komorze bezechowej Katedry Systemów Multimedialnych Politechniki Gdańskiej. Jako symulator ludzkiego torsu i głowy został użyty manekin Brüel & Kjær typ 4128-C. Sygnał rejestrowały dwa mikrofony douszne FG-23629. Posiadały one swój bateryjny układ zasilający, a sygnał był transmitowany przez przewody XLR podłączone do rejestratora ZOOM F4 MultiTrack Field Recorder. Z niego również był transmitowany sygnał wzorcowy do wyjścia słuchawkowego. Sygnałem wzorcowym był sygnał świergotowy, przemiatający (ang. chirp, sine sweep). Rejestrator został również podłączony do komputera przenośnego w trybie interfejsu audio. Na komputerze przenośnym wyzwalano jedną pętlę pomiarową sterującą rejestratorem oraz zapisywano wyniki pomiaru.

Skrypt został zaprojektowany i uruchamiany w programie Pure Data. Wykonano 10 pomiarów dla każdego modelu słuchawek, które poddano dalszemu przetwarzaniu.

Mikrofony pomiarowe zostały umieszczone przy wejściu kanału usznego manekina za pomocą okrągłych pianek. W ten sposób zapewniono stabilność mikrofonom i zamknięto kanał. Przy każdym pojedynczym pomiarze zdejmowano i zakładano słuchawki ponownie. Sygnał wzorcowy nie był emitowany w tym samym czasie dla obu słuchawek – najpierw mierzono lewy kanał, a następnie w odstępie około 0,5 sekundy – prawy kanał.

Po skompletowaniu bazy sygnałów odpowiedzi impulsowych słuchawek przystąpiono do przetwarzania sygnałów. Sygnały każdej próbki sygnałowej przetworzono w programie Audacity, usuwając zbędne opóźnienia. Następnie przesunięto oba wycięte fragmenty sygnałów, aby rozpoczęły się w tym samym czasie i trwały tyle samo, co nadany sygnał wzorcowy, czyli dwie sekundy.

Do uzyskania funkcji HpTF wykorzystano ogólnodostępne narzędzie „AKtools” [6], zaimplemento-

wane w programie MATLAB. Narzędzie to jest zbiorem skryptów umożliwiających przetwarzanie sygnałów w dziedzinie nagrań binauralnych. Ponadto, dla każdego narzędzia autorzy przygotowali skrypty demonstracyjne, dodatkowo objaśniające działanie kodu.

Ze zbioru narzędzi AKTools wytypowano dwa właściwe do przetwarzania HpTF. Pierwszym narzędziem jest AKdeconv.m, służącym do rozplotu sygnału. Drugim skryptem jest AKregulatedInversion.m, obliczającym właściwe funkcje HpTF. Co warto podkreślić, skrypt ten zawiera różne metody regularyzacji danych wejściowych różniące się złożonością i jakością. Ponadto, skrypt ten umożliwia uśrednianie funkcji przenoszenia oraz zapisywanie wygenerowanych filtrów. W kodzie zaimplementowano 6 różnych sposobów na otrzymanie pożądanych filtrów; autorzy zarekomendowali dwa z nich, które w przypadku słuchawek dają najlepsze rezultaty. Dodatkowo skrypt umożliwia implementację różnych filtrów w celu polepszania jakości uzyskiwanych HpTFów.

W niniejszej publikacji przedstawiono szczegółowe informacje na temat utworzonej bazy. Opisano jej konstrukcję oraz sposób obróbki plików sygnałowych. Dodatkowo przeprowadzono szczegółowe analizy wybranych próbek sygnałów fonicznych zawartych w bazie wraz z interpretacją.

Ponadto zaprezentowano podobieństwa i różnice dla trzech modeli przebadanych słuchawek. W podsumowaniu pracy podano wnioski, wynikające z uzyskanych wyników wraz ze szczególnym opisem aspektów, które okazały się kluczowe w przeprowadzonych pomiarach. Dodatkowo wskazano dalsze kierunki rozwoju niniejszej bazy. Wskazano również narzędzia do pracy z sygnałami HpTF zamieszczonymi w niniejszej bazie odpowiedzi impulsowych.

ŹRÓDŁA:

[1] "ARI HpIR Database," Online. Available: https://sofacoustics.org/data/headphones/ari/ [Accessed Sep. 25, 2024],
[2] B. Boren, M. Geronazzo, P. Majdak, E. Choueiri, "Phona: a public dataset of measured headphone transfer functions," in 137th Audio Engineering Society Convention, Los Angeles, CA, USA, October 9-12, 2014,
[3] F. Brinkmann, A. Lindau , S. Weinzierl, G. Geissler, S. van de Par, M. Müller-Trapet, R. Obdam,
M. Vorländer; "The FABIAN head-related transfer function data base," Online. Available: http://dx.doi.org/10.14279/depositonce-5718 [Accessed Sep. 25, 2024], 2017,
[4] B. Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," in DAGA Fortschritte der Akustik, Meran, Italy, March, 2013,
[5] "Virtuoso," Online. Available: https://apl-hud.com/product/virtuoso/ [Accessed Sep. 25, 2024],
[6] F. Brinkmann, S. Weinzierl, "AKtools – An Open Software Toolbox for Signal Acquisition, Processing, and Inspection in Acoustics," in 142nd Audio Engineering Society Convention, Berlin, Germany, May 20-23, 2017.

Paper ID: **7**

THE ANGULAR POSITION OF A SOUND SOURCE AFFECTS THE PERCEPTION OF SCARINESS

Antonina Stefanowska, Sławomir K. Zieliński
Department of Digital Media and Computer Graphics, Faculty of Computer Science, Białystok University of Technology
Corresponding author: **Sławomir K. Zieliński**, *s.zielinski@pb.edu.pl*

Keywords: spatial sound, emotion, scariness

INTRODUCTION

As audio technology advances and spatial sound reproduction devices become more prevalent in households, sound design with an emphasis on intensified experiences may prove to be a valuable area of study. Sight is considered the most developed sense among humans and the main source of the majority of sensory information received by our brains [4]. Consequently, it is not surprising that primal survival instincts encourage us to keep any potential threat in our field of view. The field of view for humans, including peripheral vision, has an approximate range of <-60°, 60°> around the fixation point [5]. Given the relationship between visibility of potential threats and the feeling of fear, it can be hypothesized that listening to a scary sound coming from outside this angular range increases the perceived scariness of the sound.

The existing literature provides evidence that the sound source localization may influence the intensity of perceived and/or experienced emotions, scariness included [3]. For instance, the angular location outside the listeners' field of view resulted in heightened arousal (intensity) and lowered valence (positivity) ratings [1], [2]. The study on scariness demonstrated that a sound outside of listeners' view that is hard to localize tends to be perceived as scarier than otherwise [3]. This research aims to expand knowledge on this topic further by conducting experiments performed on eight angular settings of sound source position and evaluating its impact on listeners' perception of scariness.

PROCEDURE

Subjective listening tests were conducted in the spatial sound laboratory of the Bialystok University of Technology. A total of thirty-six students participated in the study. The listeners were requested to assess each sound sample based on their perception of its scariness. The audio was reproduced by a system of eight active loudspeakers arranged in a circle with a radius of two meters around the listener. The angular distance between adjacent loudspeakers was 45°. Three recordings from the IADS-E dataset [6] were selected for their relatively high ratings of fear. They included glass shattering, door banging, and hog growling. A total of 24 unique sound samples were employed in the tests (3 recordings × 8 angular variants). During the listening tests, sounds were played sequentially from the eight loudspeakers arranged around the listener, with each loudspeaker corresponding to a specific angular variant. The playback of samples was controlled by the participants using a custom-developed application with a graphical interface. Every sound was assessed on a 7-point scale ranging from "*Not scary at all"* to "*Extremely scary*."

RESULTS

A three-way ANOVA was conducted on the collected data. The effects of the angle, recording type, and participants were found to be statistically significant ($p < 0.01$). Subsequently, Tukey's HSD post-hoc pairwise tests showed that the -90° angular setting was assessed as scarier than both 0° ($p = 0.015$) and 45° ($p = 0.012$) settings. The *t*-test for independent samples was used to compare the ratings between the angular positions withing the field of view and those outside it. The results indicate that the scariness ratings within the field of view were statistically significantly lower than those outside it ($p = 0.0015$).

Overall, the obtained results indicate that the angular position of a sound source affects the perception of scariness. Specifically, sound sources positioned outside the field of view were perceived as more scary than those located within the field of view. This aligns with the initial hypothesis and results from similar studies [1], [2], [3]. Further studies could benefit from a more comprehensive examination of a wider range of emotions and the analysis of physiological responses acquired from listeners via wearable sensors.

REFERENCES

[1] ASUTAY E., VÄSTFJÄLL D., Attentional and Emotional Prioritization of the Sounds Occurring Outside the Visual Field, Emotion, 15(3): 281-286, 2015. doi: 10.1037/emo0000045

[2] DROSSOS K., FLOROS A., GIANNAKOULOPOULOS A., KANELLOPOULOS N., Investigating the Impact of Sound Angular Position on the Listener Affective State, IEEE Transactions on Affective Computing, 6(1): 27–42, 2015. doi: 10.1109/TAFFC.2015.2392768

[3] EKMAN I., KAJASTILA R., Localization Cues Affect Emotional Judgments – Results from a User Study on Scary Sound, 35th Audio Engineering Society International Conference: Audio for Games, London, United Kingdom, 166–171, 2009. url: http://www.aes.org/e-lib/browse.cfm?elib=15177

[4] HUTMACHER F., Why Is There So Much More Research on Vision Than Any Other Sensory Modality?, Frontiers in Psychology, 10: 2246, 2019. doi: 10.3389/fpsyg.2019.02246

[5] SARDEGNA J., SHELLY S., RUTZEN A. R., Scott M. S., The Encyclopedia of Blindness and Vision Impairment, Infobase Publishing, 2002.

[6] YANG W., MAKITA K., NAKAO T., KANAYAMA N., MACHIZAWA M.G., SASAOKA T., SUGATA A., KOBAYASHI R., HIRAMOTO R., YAMAWAKI S., IWANAGA M., MIYATANI M., Affective auditory stimulus database: An expanded version of the International Affective Digitized Sounds (IADS-E). in: Behavior Research Methods, 1–15, 2018.

Paper ID: **8**

MOŻLIWOŚCI WYKORZYSTANIA POMIARÓW AKUSTYCZNYCH EMISJI ULTRADŹWIĘKOWEJ ROŚLIN NA POTRZEBY SYSTEMÓW WSPOMAGANIA DECYZJI W OGRODNICTWIE

Bartłomiej Chojnacki [1], Klara Chojnacka [1], Piotr Książek [2], Janusz Mazurek [3], Wiktoria Potoniec [1], Piotr Chohura [3]
1 Katedra Mechaniki i Wibroakustyki, Wydział Inżynierii Mechanicznej i Robotyki, Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
2 Katedra Akustyki, Mutlimediów i Przetwarzania Sygnałów, Wydział Elektroniki, Fotoniki i Mikrosystemów, Politechnika Wrocławska
3 Katedra Ogrodnictwa, Wydział Przyrodniczo-Technologiczny, Uniwersytet Przyrodniczy we Wrocławiu
Autor korespondencyjny: **Bartłomiej Chojnacki**, *bchojnacki@agh.edu.pl*

Słowa kluczowe: fitoakustyka, emisja ultradźwiękowa roślin, ultradźwięki, przemysł 4.0

1. WPROWADZENIE

Fitoakustyka (ang. *phytoacoustic*) jest dziedziną akustyki wydzieloną już kilkadziesiąt lat temu, jednak dopiero na skutek niedawnych przełomów w tej dziedzinie zaistniały możliwości zastosowania wiedzy z tego zakresu w przemyśle ogrodniczym [1]. Większość dotychczas prowadzonych badań skupiała się na rejestracji dźwięków wydawanych przez rośliny w fazie wzrostu i rozwoju [2][3]. W wypadku stosowania pobudzeń akustycznych, zazwyczaj były to tony proste [4] lub pobudzenie sygnałami muzycznymi lub szumowymi [5]. Odkrycie ultradźwiękowej emisji akustycznej (UEA) od roślin i wskazanie na jej informacyjny charakter [6] umożliwiło prowadzenie nowych badań, zorientowanych na dostarczenie metod pomiarowych na potrzeby m. in. systemów wspomagania decyzji w ogrodnictwie. Rejestracja UEA od roślin pozwala na zdiagnozowanie stanu rośliny, m. in. w zakresie odwodnienia, ataku szkodnika lub wystąpienia szkód mechanicznych. W niniejszym referacie zaprezentowanie zostanie koncepcja badań fitoakustycznych prowadzonych w Akademii Górniczo-Hutniczej w Krakowie oraz wyniki badań pilotażowych pomiarów pomidorów (*Solanum lycopersicum*) w komorze bezechowej AGH.

2. METODA BADAWCZA

W ramach badań pilotażowych UEA wykonano pomiary sadzonki pomidora w komorze bezechowej AGH w Krakowie. Wykorzystując zestaw mikrofonów GRAS 46BE ¼” (zakres częstotliwości do 80 kHz), badano kierunkowość emisji dźwięku od sadzonki, rozstawiając mikrofony w okręgu o promieniu 30 cm. Mikrofony umieszczone były co 45 stopni na całym obwodzie okręgu. Dodatkowo, wykorzystano dwa mikrofony AviSoft CM16 o maksymalnym zakresie częstotliwościowym do 300 kHz i kartę dźwiękową AviSoft 416H o częstotliwości próbkowania 750 kHz . Fotografię stanowiska badawczego pokazano na rys. 1. Badanie polegało na pomiarze sadzonki pomidora w warunkach umożliwiających normalną wegetację rośliny, poddając ją jednak stresowi wodnemu (wstrzymano podlewanie). Roślina była doświetlana zgodnie z zaleceniami przez 16 godzin na dobę.

Rys. 1. Fotografia stanowiska badawczego w komorze bezechowej do pomiaru kierunkowości ultradźwiękowej emisji akustycznej od sadzonki pomidora

Pomiar trwał 10 dni, rejestrację prowadzono w trybie ciągłym. W uzyskanym materiale odnaleziono impulsy ultradźwiękowe występujące paśmie częstotliwości od 30 do 50 kHz, co potwierdza wstępne założenia analizy UEA roślin tego typu dostępne w literaturze. Wraz ze wzrostem stresu wodnego, wzrastała liczba impulsów rejestrowanych przez aparaturę. W wypadku rośliny zdrowej było to od 4 do 10 impulsów w ciągu godziny, dla rośliny odwodnionej (podczas 7 dnia pomiaru) rejestrowano już ponad 70 impulsów na godzinę.

3. PODSUMOWANIE

Przeprowadzone badania pilotażowe potwierdziły informacyjny i użyteczny charakter UEA pochodzącej od roślin. Dalsze badania kierunkowości dźwięku oraz badania wibracji generowanych przez sadzonkę w trakcie emisji pozwolą na szczegółowe zbadanie genezy tego zjawiska i ustalenie, które elementy rośliny odpowiadają ze UEA. W przyszłości zaplanowano opracowanie metod algorytmicznych pozwalających na precyzyjną diagnostykę stanu rośliny i wykrywanie różnych zagrożeń na podstawie badania ultradźwiękowych emisji akustycznych roślin. Badania tego typu mogą posłużyć do stworzenia systemów wspomagania decyzji dla ogrodników, znacznie przyczyniając się do rozwoju nowych technologii w ogrodnictwie.

LITERATURA

[1] Ali S, Tyagi A, Park S, Bae H, Understanding the mechanobiology of phytoacoustics through molecular Lens: Mechanisms and future perspectives, J. Adv. Res., [Internet] 2023, . Available from: https://www.sciencedirect.com/science/article/pii/S2090123223003983
[2] Son J-S, Jang S, Mathevon N, Ryu C-M, Is plant acoustic communication fact or fiction?, New Phytol., [Internet] 242 (3)5, 2024, 1876–80,. Available from: https://doi.org/10.1111/nph.19648
[3] Allievi S, Arru L, Forti L, A tuning point in plant acoustics investigation, Plant Signal. Behav., [Internet] 16 (3)8, 2021, . Available from: https://doi.org/10.1080/15592324.2021.1919836
[4] Jeong MJ, Cho J Il, Park SH, Kim KH, Lee SK, Kwon TR, et al., Sound frequencies induce drought tolerance in rice plant, Pakistan J. Bot., 46 (3)6, 2014, 2015–20,.
[5] Shivanna KR, Phytoacoustics - Plants can perceive ambient sound and respond, J. Indian Bot. Soc., 102 (3)1, 2022, 1–5,.
[6] Khait I, Lewin-Epstein O, Sharon R, Saban K, Goldstein R, Anikster Y, et al., Sounds emitted by plants under stress are airborne and informative, Cell, [Internet] 186 (3)7, 2023, 1328-1336.e10,. Available from: https://doi.org/10.1016/j.cell.2023.03.009

Paper ID: **9**

LOCALIZATION ACCURACY OF THE SELECTED METHODS USED FOR SPATIAL AUDIO RENDERING IN VIRTUAL-REALITY SYSTEMS

Ihar Balyka, Sławomir K. Zieliński
Department of Digital Media and Computer Graphics, Faculty of Computer Science, Białystok University of Technology
Corresponding author: **Sławomir K. Zieliński**, *s.zielinski@pb.edu.pl*

Keywords: Binaural audio rendering, virtual reality

1. INTRODUCTION

Binaural technologies are nowadays commonly used to render spatial audio in virtual reality systems, significantly enhancing immersion in simulators, games, films, and music. While these technologies have been greatly improved over the past decades, even the state-of-the-art techniques still exhibit some deficiencies in terms of their localization accuracy or timbral fidelity [4]. The primary objective of this study was to compare the five selected binaural rendering techniques commonly used in virtual reality systems. The technologies under scrutiny included Unreal Engine Audio, FMOD, Wwise, Steam Audio, and Resonance Audio. The goals of the study were twofold: to verify the utility of the head tracker in the context of modern spatial audio rendering software (middleware) used for computer games, and to benchmark (compare) selected versions of state-of-the-art spatial audio rendering software used for games in terms of localization accuracy.

During the preparation of this work, other comparisons reported in the literature were examined. However, many of them were outdated. For instance, a study by Catalano from 2011 [3] showed no significant differences and generally poor performance in spatial audio for FMOD and Wwise. Steam Audio was separately studied in 2019 [2], as well as Resonance Audio [5]. Since then, technologies have advanced significantly [4], highlighting the need for more recent research. This work employs state-ofthe-art spatial audio technologies to ensure comprehensive and up-to-date comparisons.

2. EXPERIMENTS AND RESULTS

Two experiments were conducted: one to investigate the impact of the head tracker, and another to examine the differences between the chosen technologies. The experiments involved listening tests undertaken in headphones in an acoustically treated laboratory. During the course of the experiments, the listeners were presented with a graphical user interface depicting eight sound sources, which were positioned equidistantly in the horizontal plane around the virtual head. The participants were instructed to listen to the stimulus (a pop music excerpt) and to indicate the perceived angle of sound incidence using a computer mouse. The apparatus consisted of Sennheiser HD 215 headphones driven by a Lexicon Alpha interface. A head-tracker was employed to enable dynamic three-degrees-of-freedom audio rendering. A custom-developed application was utilized to facilitate the listening tests. In the first experiment, a total of 736 responses were collected from 14 users. A post-screening procedure was employed to remove the data from unreliable listeners. According to the obtained results, the mean localization error with the head tracker was equal 25.7°, compared to 49.6° without it. The findings confirm

the significant impact of the head tracker on improving the localization accuracy [1].

In the second experiment, a panel of 15 participants were involved in the listening tests, yielding 300 responses in total. According to the results, there were notable differences in the localization accuracy between the technologies. The Steam Audio proved to be the most accurate technology with a mean localization error of 14.8°, outperforming FMOD by approximately 5°. In contrast, the Resonance Audio exhibited mediocre performance, with a localization error of 23.9°. Notably, both Wwise and Unreal Engine Audio showed the poorest performance, with localization errors of 38.0° and 43.6°, respectively.

2. SUMMARY

The experiments revealed that the state-of-the-art binaural rendering technologies exhibit pronounced differences in terms of the localization accuracy. According to the obtained results, Steam Audio proved to be the most accurate technique, followed by FMOD and Resonance Audio. Wwise and Unreal Engine Audio showed comparatively lower accuracy. Moreover, our experiments confirmed that the use of a head-tracker is indispensable in binaural rendering audio systems as it helps to reduce the mean localization error by approximately 50%. To further enhance the localization accuracy of the binaural audio rendering techniques, it is proposed that the differences in pinna, headphones, and their coupling effects are taken into account in the individualization procedure performed by a listener.

REFERENCES

[1] BEGAULT D., WENZEL E., Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source, Journal of the Audio Engineering Society, Vol. 49, No. 10, 2001, 904-916.
[2] BOUCHARA T., BARA T.-G., WEISS P.-L., GUILBERT A., Influence of vision on short-term sound localization training with non-individualized HRTF, In Proc. of the EAA Spatial Audio Signal Processing Symposium, Paris, France, Sep. 2019, 55-60
[3] CATALANO G., Virtual Reality in Interactive Environments: A Comparative Analysis of Spatial Audio Engines, S.A.E. Institute Oxford, 2011.
[4] PATERSON J., LEE H. (Eds.), 3D Audio, Routledge, London, 2022.
[5] POURU L., The Parameters of Realistic Spatial Audio: An Experiment with Directivity and Immersion, Bachelor's thesis, Turku University of Applied Sciences, 2019.

Paper ID: **10**

SYNTEZA IMMERSYJNYCH WRAŻEŃ W ODSŁUCHU SŁUCHAWKOWYM Z WYKORZYSTANIEM KIERUNKOWYCH ODPOWIEDZI IMPULSOWYCH POMIESZCZENIA

Witold Mickiewicz, Kaja Kosmenda
Katedra Inżynierii Systemów, Sygnałów i Elektroniki, Wydział Elektroczny,
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie
Autor korespondencyjny: **Witold Mickiewicz**, *witold.mickiewicz@zut.edu.pl*

Słowa kluczowe: dźwięk immersyjny, odsłuch słuchawkowy, konwersja binauralna

1. WPROWADZENIE

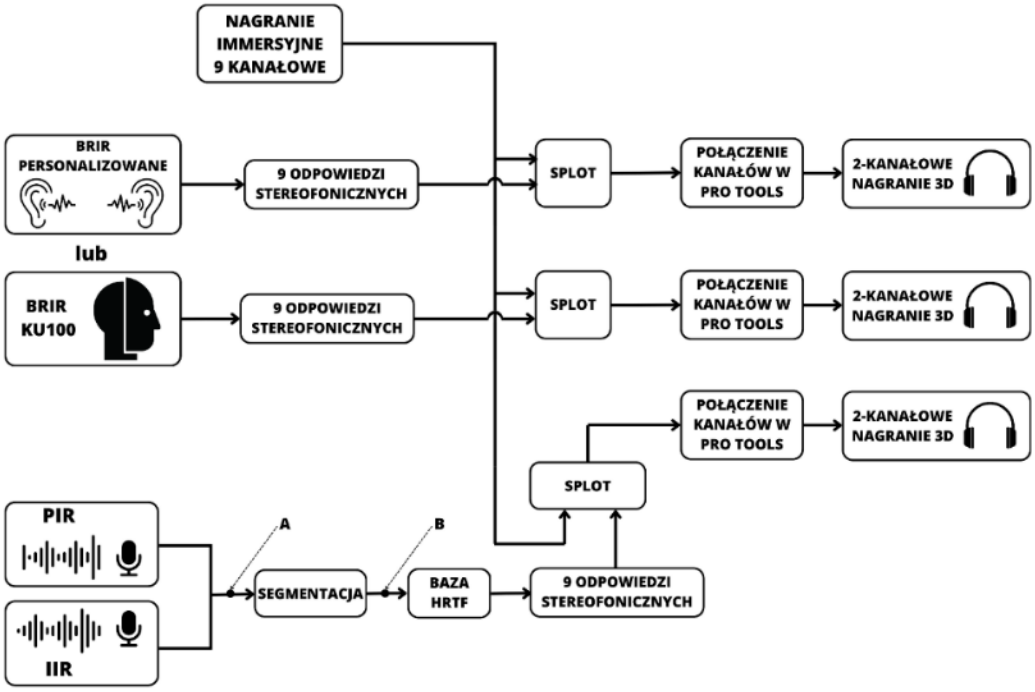
Niniejsza praca koncentruje się na technice przetwarzania wielokanałowego dźwięku w taki sposób, aby przy odsłuchu słuchawkowym zachować immersyjne wrażenia dźwiękowe obecne podczas słuchania nagrań w dobrej jakości pomieszczeniu odsłuchowym wyposażonym w system głośnikowy 7.1.4., który gwarantuje wytworzenie wrażeń immersyjnych na dobrym poziomie. Dzięki takiemu podejściu możliwe było dokonanie szczegółowej i wiarygodnej analizy porównawczej wrażeń słuchaczy podczas odtwarzania nagrań z głośników z tymi, które powstają przy odbiorze przetworzonych nagrań przez słuchawki. Analiza wyników badań pozwala na walidację zaprojektowanej techniki przetwarzania dźwięku, oferując wgląd w potencjalne kierunki jej dalszego rozwoju i zastosowania. W referacie zaprezentowano algorytm do przetwarzania immersyjnego nagrania wielokanałowego do odsłuchu słuchawkowego bazujący na fuzji danych pomiarowych z dwóch czujników: mikrofonu ciśnieniowego oraz natężeniowej sondy PU. Za ich pomocą rejestrowano charakterystykę ciśnieniową i kierunkową pomieszczenia odsłuchowego. Odpowiednia segmentacja tych danych uzupełniona o dane z uśrednionych baz HRTF umożliwiła syntezę zestawu odpowiedzi impulsowych użytych następnie w procesorze splotowym do stworzenia nagrania w wersji immersyjnej.

2. PRZEPROWADZONE BADANIA

W ramach badań porównano efekty działania algorytmu segmentacji odpowiedzi impulsowej z wykorzystaniem ogólnodostępnych baz uśrednionych odpowiedzi HRTF z klasycznym podejściem wykorzystującym zindywidualizowane binauralne odpowiedzi impulsowe pomieszczenia. Zakres przeprowadzonych badań zilustrowano na rys.1.

Porównawczy materiał dźwiękowy zarejestrowano w nowoutworzonym laboratorium dźwięku immersyjnego. Z wykorzystaniem sztucznej głowy rejestrowano uśrednione odpowiedzi impulsowe pomieszczenia, za pomocą binauralnych mikrofonów DPA zarejestrowano zestaw zindywidualizowanych odpowiedzi impulsowych pomieszczeń, a z wykorzystaniem dookólnego mikrofonu Schoeps MK5 oraz sondy natężeniowej Microflown. Wspomniane zaplecze badawcze pokazano na rys.2.

Efekty działania prezentowanego algorytmu, mimo użycia danych niezindywidualizowanych są obiecujące i skłaniają do refleksji nad rolą filtracji HRTF poszczególnych odbić dźwięku w pomieszczeniu w wywoływaniu wrażenia zanurzenia w dźwięku. Można postawić tezę, że w przypadku symulowania naturalnego środowiska odsłuchowego spada rola zindywidualizowania HRTF na korzyść ich poprawnego dynamicznego przełączania w zależności od kierunku docierania kolejnych odbić.



Rys.1. Zakres subiektywnych badań przeprowadzonych w ramach prezentowanej pracy.

Rys.2. Sztuczna głowa, mikrofony binauralne, sonda natężeniowa i laboratorium dźwięku immersyjnego z systemem głośnikowym 7.1.4.

LITERATURA

[1] MICKIEWICZ W., KOSMENDA M., Spatialization of sound recordings using intensity impulse responses, 2023 MMAR, Międzyzdroje, Poland, 2023, pp. 264-268
[2] PFANZAGL-CARDONE E., The Art and Science of 3D Audio Recording, Stringer, 2023
[3] Ed. ROGINSKA A., GELUSO P., Immersive Sound. The Art and Science of Binaural and Multi-Channel Audio, Taylor-Francis, 2018

Paper ID: **11**

REKONSTRUKCJA PÓŹNEJ CZĘŚCI ODPOWIEDZI IMPULSOWEJ POMIESZCZENIA W PROCESIE RENDEROWANIA PRZESTRZENNEGO SYGNAŁU AUDIO W CZASIE RZECZYWISTYM

Maria Peńsko, Piotr Cenda,
SoftServe Poland Sp. z o. o.
Autorzy korespondencyjni: **Maria Peńsko**, *mpens@softserveinc.com*,
Piotr Cenda *pcend@softserveinc.com*

Słowa kluczowe: odpowiedź impulsowa pomieszczenia, wirtualna rzeczywistość, audio plugin, DNN

W celu zapewnienia w pełni immersyjnego doświadczenia VR/AR konieczne jest zapewnienie realistycznej akustyki generowanej przestrzeni wirtualnej. Można to osiągnąć poprzez splot sygnału audio ze spodziewaną odpowiedzią impulsową pomieszczenia (RIR). Uzyskanie wiarygodnego efektu jest możliwe tylko przy zachowaniu wysokiej dbałości o dokładną rekonstrukcję RIR dla każdej wymaganej pary pozycji źródła dźwięku i słuchacza we wnętrzu. Takie rozwiązanie jest trudne do wdrożenia w czasie rzeczywistym.

W poprzedniej pracy [1] przedstawiliśmy metodę opartą na DNN do szybkiej predykcji wczesnej części RIR dla dowolnej lokalizacji źródła dźwięku i słuchacza w wirtualnym pomieszczeniu. Niniejsza praca uzupełnia poprzednią i koncentruje się na szybkiej rekonstrukcji późnej części RIR. Przedstawiamy naszą adaptację opartego na DNN rozwiązania odtwarzającego późną część RIR na podstawie jej wczesnej części. Omawiamy wprowadzone zmiany oraz analizujemy uzyskane wyniki.

Zaprezentowane rozwiązanie umożliwia przewidywanie realistycznej, pełnej długości odpowiedzi impulsowej pomieszczenia RIR dla dowolnej lokalizacji źródła dźwięku i słuchacza w czasie rzeczywistym.

1. WPROWADZENIE

Pierwsze 80 ms RIR obejmuje dochodzący do słuchacza dźwięk bezpośredni oraz grupę wczesnych odbić. Dalsza część RIR to późne odbicia, tzw. „ogon" pogłosowy. W naszej pracy celem jest uzupełnienie wygenerowanego wczesnego fragmentu RIR zrekonstruowanym „ogonem" pogłosowym.

Poszukując rozwiązania tego problemu w literaturze natrafiliśmy na tylko jeden opis rozwiązania.

1.1 DECOR

DECOR (Deep Exponential Completion Of Room impulse responses) [2] bazuje na architekturze FiNS (Filter Noise Shaping) [3]. Enkoder znajduje reprezentację dla wczesnej części RIR, która jest następnie wykorzystywana przez Dekoder do obliczenia krzywych wykładniczego zaniku (EDC, Exponential Decay Curve) energii akustycznej. Podlegający treningowi razem z Enkoderem i Dekoderem bank filtrów kształtuje grupę sygnałów szumu wąskopasmowego. Obliczone krzywe wykładniczego zaniku energii akustycznej są kolejno nakładane na pasma szumu, a po zsumowaniu tworzą brakujący, późny fragment RIR.

2. ROZWIĄZANIE PROBLEMU

2.1. DATASET

Korzystamy ze zbioru 10 000 odpowiedzi impulsowych wygenerowanego przy użyciu Treble SDK [4]. Odpowiedzi impulsowe charakteryzuje różny czas pogłosu (od 0,2 do 1,3 sek.) i struktura wczesnych odbić. Do wejścia modelu wprowadzany jest wczesny fragment RIR – 80 ms, co przy częstotliwości próbkowania 32 kHz odpowiada 2 560 próbkom sygnału. Na wyjściu modelu otrzymujemy 45 440 próbek „ogona" pogłosowego, co razem daje 48 000 próbek – 1,5 sek. odpowiedzi impulsowej. Odpowiedzi impulsowe krótsze niż 1,5 sek. zostały uzupełnione szumem szerokopasmowym o poziomie RMS -100 dB. W odróżnieniu od autorów [2] nie ucinaliśmy początku RIR, zostawiając poprzedzającą ją ciszę jako informację dotyczącą odległości słuchacza od źródła dźwięku.

2.2. FUNKCJA KOSZTU

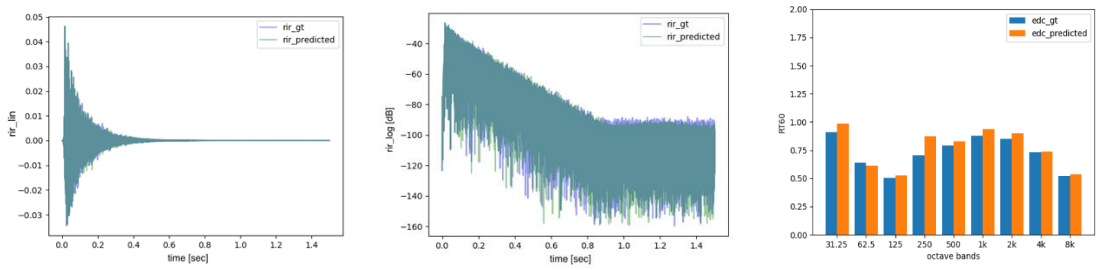
Podobnie jak autorzy [2] używamy funkcji kosztu MSTFT (Multiresolution Short Time Fourier Transform) [5]. Do jej wartości dodajemy składową związaną ze średnim błędem czasu pogłosu liczonego w pasmach oktawowych (oktawy o częstotliwościach od 31,5 do 8 000 Hz).

2.3. PARAMETRY TRENINGU

Większość parametrów treningu odpowiada opisanym w [2]. Te, którymi manipulujemy wpływają głównie na szybkość uczenia modelu, nie jego efekt. Parametry danych wejściowych i wyjściowych podyktowane są wymaganiami projektowymi.

3. WYNIKI

Poniższe wykresy przedstawiają przykład zrekonstruowanej odpowiedzi impulsowej.



Rys. 1. Oryginalna (gt_rir) i zrekonstruowana (predicted_rir) odpowiedź impulsowa w skali liniowej i logarytmicznej oraz wyznaczone z nich czasy pogłosu w pasmach oktawowych.

Wytrenowany przez nas model osiągnął podczas walidacji wartość MSTFT 1.103. Jest to wynik zbliżony do DECOR [2] – 1.073 (przy długości rekonstruowanej RIR 1.0 sek.). Osiągnięta przez nas dokładność rekonstrukcji wyrażona błędem RT60 w oktawie o częstotliwości środkowej 1 kHz wyniosła 0,051 sek.

Z wstępnie przeprowadzonych testów subiektywnych wynika, że w większości przypadków słuchacz nie jest w stanie odróżnić sygnałów mowy splecionej z oryginalną i zrekonstruowaną RIR.

4. DALSZE PLANY

Dalsze plany pracy obejmują: określenie zdolności generalizowania modelu przy użyciu zewnętrznych zbiorów odpowiedzi impulsowych, pełną ocenę subiektywną, rekonstrukcję ambisonicznych odpowiedzi impulsowych, eksperymenty ze zmodyfikowaną funkcją kosztu.

LITERATURA

[1] Cenda P., Januszkiewicz L., Wasilewski J., *ML-DRIVEN SPATIAL AUDIO RENDERING FOR ENCLOSED VIRTUAL SPACES* 19th International Symposium on Sound Engineering and Tonmeistering, Warszawa 12-14 October 2023.
[2] Lin J., Gotz G., Schlecht S. J., *Deep Room Impulse Response Completion*, 2024.
[3] Steinmetz, Christian J. and Ithapu, Vamsi Krishna and Calamia, Paul, *Filtered Noise Shaping for Time Domain Room Impulse Response Estimation From Reverberant Speech*, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2021.
[4] Treble SDK https://www.treble.tech/software-development-kit.
[5] R. Yamamoto, E. Song, and J.-M. Kim, *Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram*, ICASSP, 2020.

Paper ID: **12**

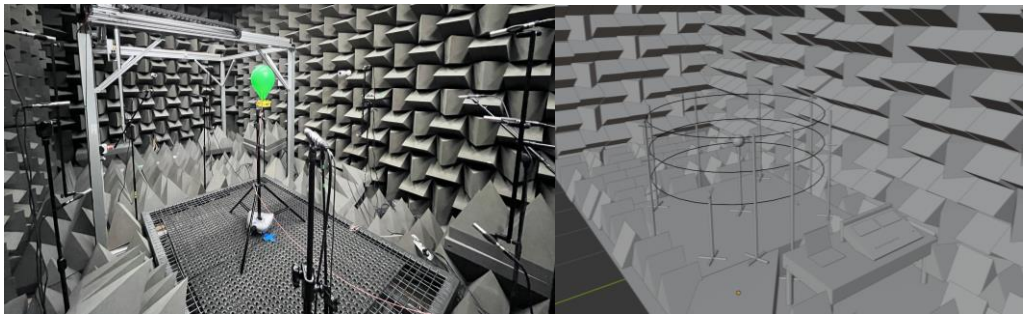
EKSPERYMENTALNE ŹRÓDŁA WSZECHKIERUNKOWE NA POTRZEBY REJESTRACJI ODPOWIEDZI IMPULSOWYCH POMIESZCZEŃ

Maciej Siłkowski
Katedra Systemów Multimedialnych, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska,
Autor korespondencyjny: **Maciej Siłkowski**, *maciek@serg.pl*

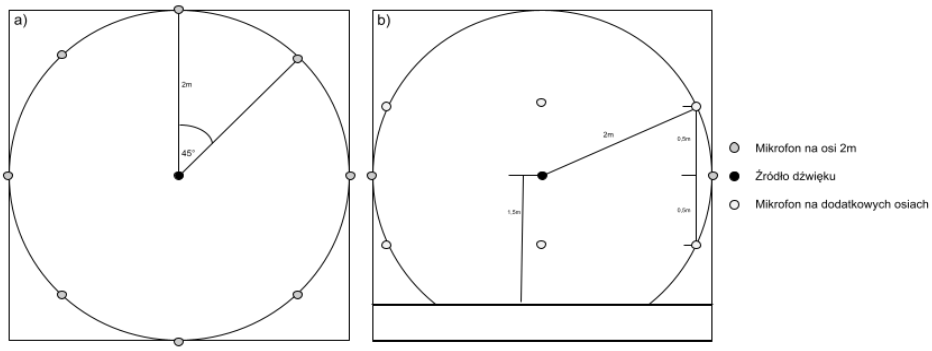
Słowa kluczowe: odpowiedź impulsowa, akustyka wnętrz, źródła wszechkierunkowe

W niniejszej pracy porównano różne eksperymentalne źródła wszechkierunkowe, które mają posłużyć jako tańsza i bardziej dostępna alternatywa dla 12-ściennych głośników klasycznie wykorzystywanych w pomiarach odpowiedzi impulsowej pomieszczeń. Porównane zostały między innymi odgłosy pękającego balonu, wydźwięk klaskania, wystrzał z rewolweru oraz układy głośnikowe.

Przygotowano stanowisko pomiarowe składające się z szesnastu mikrofonów. Rozmieszczono je w formie wycinka sfery o promieniu 2 m na trzech wysokościach. Środek sfery zajmowało testowane źródło impulsu. Układ pomiarowy przedstawiono na rys 1.



Rys. 1. Układ pomiarowy zawarty w komorze bezechowej na przykładzie pomiaru balonu oraz render z programu blender



Rys. 2. Plan rozstawienia mikrofonów widziany a) z góry oraz b) z boku

Rejestracje dźwięku przeprowadzano z próbkowaniem 48kHz/24 bit. W przypadku jednego z pobudzeń, zwiększono próbkowanie do 96kHz. W przypadku niektórych źródeł dźwięku, na przykład balonu, równocześnie dokonywano rejestracji kamerą szybkoklatkową, co pozwoliło na dokładne śledzenie puknktu przebicia balona i proces jego rozpadu. Przebadano kilka rozmiarów i kształtów balonów, o średnicy 25 i 30 cm oraz w kształcie serca. Dodatkowo, przebijana dokonywano za pomocą specjalnie opracowanego mechanizmu wykluczającego obecność człowieka, jak i powtórzono pomiary dla odpowiednio kobiety i mężczyzny. Klaskanie wykonywane przez różne osoby po dwa układy dłoni (rys. 3). W sumie zbadano X osób (y mężczyzm, y kobiet). Wystrzał z rewolweru również badany był z obecnością człowieka, jak i z układem pozwalającym na pobudzenie wystrzału zpoza strefą pomiarową. Na koniec sprawdzono kierunkowość macierzy ośmiu głośników ustawionych w dwóch różnych konfiguracjach.

Rys. 3. Układ rąk podczas klaskania oraz przykład układu głośnikowego

Wstępne wyniki otrzymane z miernika SPL, prezentują zadowalający poziom dźwięku względem szumu tła oraz dynamikę pozwalającą na wykorzystanie przy pomiarach odpowiedzi impulsowych.

LITERATURA

[1] Papadakis, N.M.; Stavroulakis, G.E. Handclap for Acoustic Measurements: Optimal Application and Limitations. *Acoustics* 2020, *2*, 224-245.
[2] Lamothe, Reina & Bradley, J.. (1985). Acoustical characteristics of guns as impulse sources. 13.
[3] Pätynen J, Katz BF, Lokki T. Investigations on the balloon as an impulse source. J Acoust Soc Am. 2011 Jan;129(1):EL27-33.

Paper ID: **13**

PROJEKT I IMPLEMENTACJA APLIKACJI VR DO PRZEPROWADZANIA TESTÓW LOKALIZACJI ŹRÓDŁA DŹWIĘKU W PRZESTRZENI

Grzegorz Rusinek, Agnieszka Pietrzak
Instytut Radioelektroniki i Technik Multimedialnych, Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska
Autor korespondencyjny: **Grzegorz Rusinek**, *grzegorz.rusinek.stud@pw.edu.pl*

W badaniach percepcji dźwięku przestrzennego wykorzystuje się obecnie często aplikacje w technologii wirtualnej rzeczywistości. W niniejszej pracy przedstawiono projekt oraz implementację aplikacji wykorzystującej trójwymiarowe środowisko wirtualne, umożliwiającej przeprowadzanie przestrzennych testów słuchowych z wykorzystaniem gogli wirtualnej rzeczywistości. Aplikacja korzysta z kontrolerów ruchu przestrzennego dołączonych do gogli w celu wskazywania położenia źródła dźwięku (metoda hand-pointing). Aplikacja umożliwia przeprowadzanie testów lokalizacji źródła dźwięku w płaszczyźnie azymutu i elewacji. Zawiera też tryb treningu, w którym po wskazaniu źródła dźwięku wyświetlana jest jego właściwa lokalizacja. Aplikacja pozwala na wyświetlanie klatki w kształcie sfery o sparametryzowanych wymiarach, pomagającej odnaleźć się użytkownikowi w przestrzeni.

Słowa kluczowe: Percepcja dźwięku przestrzennego, przestrzenne testy słuchowe, wirtualna rzeczywistość, położenie źródła dźwięku

Paper ID: **14**

OPTYMALIZACJA SPOSOBU REALIZACJI NAGRAŃ MUZYCZNYCH I TECHNIK MIKROFONOWYCH POD KĄTEM ROZSZERZENIA STEREOFONII DO SYSTEMÓW 3D.

Łukasz Kurzawski
Wydział Kompozycji, Teorii Muzyki i Reżyserii Dźwięku, Akademia Muzyczna im. Feliksa Nowowiejskiego w Bydgoszczy
Autor korespondencyjny: **Łukasz Kurzawski**, *info@recart.pl*

Trwa ewolucja technologiczna, której najbardziej ogólnym założeniem jest dalsze dostosowywanie sposobu realizacji nagrań do wysublimowanych i bardzo złożonych możliwości ludzkiej percepcji oraz poszerzenie możliwości artystycznej ekspresji twórców.

Terminem coraz częściej pojawiającym się w kontekście odbioru sztuki, również nagrań muzycznych, jest immersja. Koncepcja ta łączy gąszcz możliwości technologicznych w służbie jednego celu – rzeczywistego odczucia, realnego udziału słuchacza w prezentowanym nagraniu. Immersja pozwala na połączenie parametrów fizycznych, technicznych i akustycznych z percepcją dźwięku i psychologią muzyki. Nagrania 3D wydają się być atrakcyjnym rozwiązaniem na polu próby osiągnięcia efektu immersji.

Dominującym formatem nagrywania i odtwarzania muzyki jest obecnie stereofonia. Podstawowe nagraniowe techniki stereofoniczne, czyli AB, XY, ORTF czy M/S, pozwalają inżynierom dźwięku w zadowalający sposób realizować swoje zamierzenia artystyczne na bazie tej technologicznej. W przypadku trójwymiarowych systemów odtwarzania dźwięku sytuacja jest znacznie bardziej skomplikowana. Opisanych obecnie systemów mikrofonowych umożliwiających stworzenie takiego nagrania jest co najmniej 250 w tym: PCMA-3D, OCT-3D, 2L-Cube, Decca Cubioid, czy Hamasaki Square.

Jak optymalnie rozszerzyć tradycyjne techniki mikrofonowe przy nagrywaniu muzyki klasycznej, aby w procesie postprodukcji uzyskać zarówno nagrania stereo, jak i formaty trójwymiarowe (np. Dolby Atmos)? W ramach referatu zostanie przedstawiony autorski system mikrofonizacji z przykładami zrealizowanych nagrań chóralnych, orkiestrowych i kameralnych zarówno sesyjnych jak i koncertowych.

Słowa kluczowe: dźwięk immersyjny, nagrania muzyki poważnej, techniki mikrofonowe 3D, Dolby Atmos

Paper ID: **15**

AUTOMATYCZNA TRANSKRYPCJA DŹWIĘKÓW PIANINA

Jakub Wolski, Patryk Gawłowski
Katedra Akustyki, Mutlimediów i Przetwarzania Sygnałów, Wydział Elektroniki, Fotoniki i Mikrosystemów, Politechnika Wrocławska
Autor korespondencyjny: **Jakub Wolski**, *jakobwolski6@gmail.com*

Słowa kluczowe: Automatyczna Transkrypcja Muzyki, sieci neuronowe, głębokie uczenie, detekcja wielotonowa

1. WSTĘP

Automatyczna transkrypcja muzyki przekształca nagrania dźwiękowe na zapis nutowy. Wymaga rozpoznania instrumentu, tempa, wysokości i czasu trwania dźwięków [2]. Skupiono się wyłącznie transkrypcji pianina, dążąc do utworzenia pliku MIDI z pliku audio. Pracę oparto na artykule "An AI Approach to Automatic Natural Music Transcription" [1].

2. METODOLOGIA I IMPLEMENTACJA

Skorzystano z powszechnie dostępnej bazy danych MAPS, która posiada nagrania wielu modeli pianina w różnych formach (pojedyncze/polifoniczne dźwięki, utwory muzyczne, etc.) oraz warunkach [4].

Pierwszym krokiem przetwarzania danych był *downsampling* plików audio do częstotliwości próbkowania 16 kHz – w tym celu wykorzystano instrukcje z pakietu Pythona *librosa* [5].

Następnie wykorzystano CQT (*Constant Q Transform*) aby przekształcić sygnał do dziedziny częstotliwościowej. Jest ona alternatywą do DFT. Główną różnicą jest to, że podczas grupowania częstotliwości CQT używa stałych proporcji – wyższe tony są bliżej siebie a niższe są dalej – lepiej odwzorowując percepcję ludzkiego słuchu i ułatwiając analizę muzyki [3]. Parametr hop_length ustawiono na 512, co daje 31.25 klatek na sekundę. Przyjęto 88 klawiszy pianina (A0 do C8) – zatem CQT tworzy macierz o wymiarach (liczba klatek, 88) dla każdego pliku muzycznego.

Kolejnym etapem było grupowanie klatek, co symuluje pamięć sieci neuronowej LSTM (*Long Short-Term Memory*), gdzie każda klatka była grupowana z trzema poprzednimi i trzema kolejnymi klatkami. W rezultacie otrzymano macierz o wymiarach (liczba klatek, 88, 7). Najważniejsza klatka znajduje się na środku grupy, a pozostałe przekazują informacje o kontekście.

Ostatnim krokiem było stworzenie etykiet do procesów uczenia i walidacji. Wykorzystano pliki MIDI i pakiet Pythona *pretty-midi* do utworzenia macierzy etykiet o wymiarach (liczba klatek, 88) złożonej z 1 i 0 (1 – klawisz jest naciśnięty, 0 – klawisz nie jest naciśnięty) [6].

Wszystkie pliki audio, które zostały wyselekcjonowane z bazy danych MAPS, przetworzono w powyższy sposób oraz zachowano w pojedynczym pliku za pomocą pakietu Pythona *pickle*.

4. SIEĆ NEURONOWA

Zaimplementowano sieć neuronową, która posiada dwie warstwy konwolucyjne o różnych rozmiarach (25x5 i 5x3) po 50 filtrów, po których występują warstwy max pooling o rozmiarach 3x5, z funkcją aktywacji tangens hiperboliczny i dropoutem 0.3. Po nich są dwie w pełni połączone wartstwy o długości odpowiednio 1000 i 200 neuronów (obie z sigmoidalną funkcją aktywacji i dropoutem 0.3).

Nauczono modele na różnych zbiorach danych: M1 (pojedyncze dźwięki), M2 (pełna baza danych od zera), M3 (utwory muzyczne) oraz M4 (rozszerzenie M1 o pełną bazę). Porównanie tych modeli pozwoliło ocenić wpływ różnorodności danych na efektywność sieci.

Do oceny modelu użyto F1-Score (R - czułość, P - precyzja). Idealny wynik to 1.0, najgorszy 0.0. F1-Score jest szczególnie użyteczna przy nierównowadze klas, gdzie accuracy mogłoby być mylące.

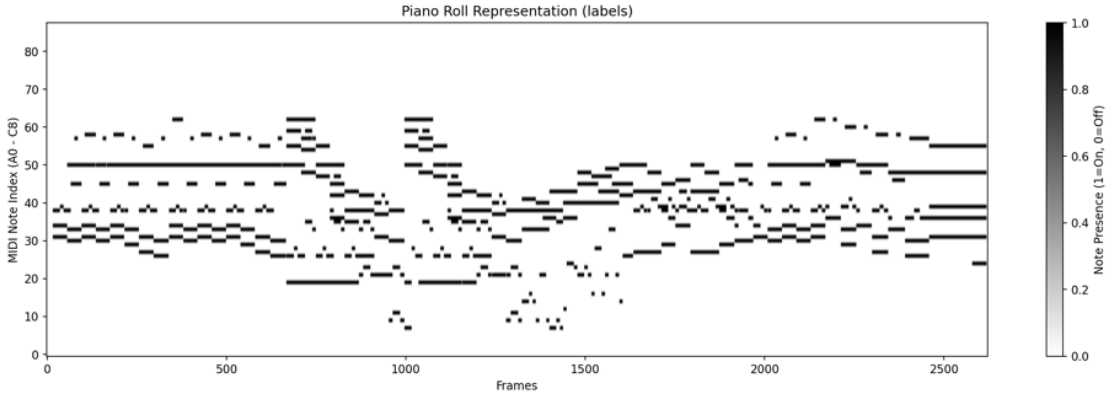
$$\text{F1-Score} = 2 * R * P / R + P \tag{1}$$

Modele uczono z podziałem danych na zbiór testowy (80%) i walidacyjny (20%). Najważniejsze parametry to learning_rate = 0.01 i batch_size = 32. Uczenie zatrzymywano, gdy F1-Score przestało się poprawiać.

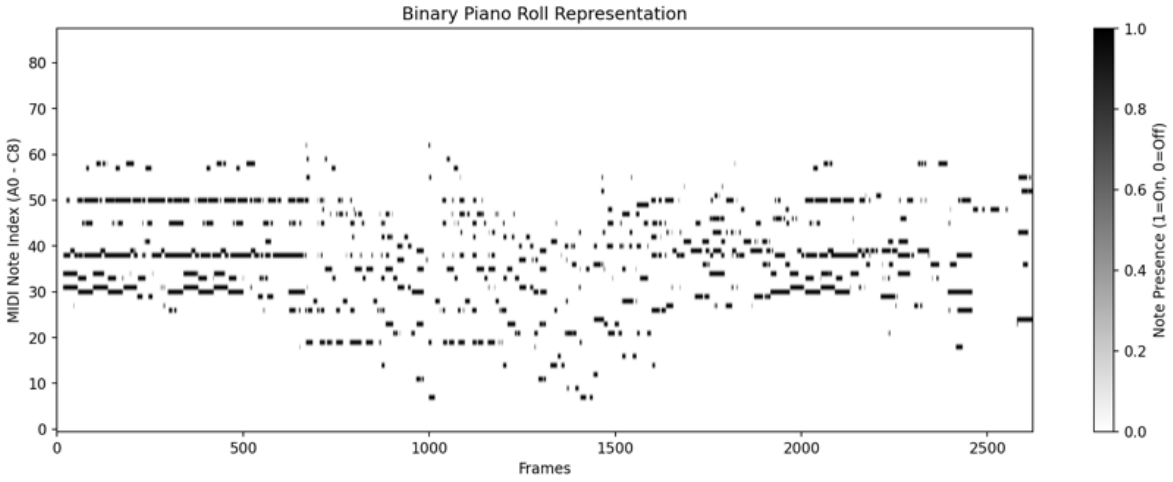
5. WYNIKI

Ostateczne wyniki F1-Score dla modeli to: M1 - 70.27%, M2 - 36.56%, M3 - 28.23%, M4 - 34.85%. Podczas analizy predykcji modeli różnych dźwięków M1 dawał najlepsze wyniki dla pojedynczych nut natomiast nie radził sobie z polifonią – z tym najlepiej radził sobie M2. Wyniki predykcji utworu muzycznego modelu M2 przedstawiono graficznie na Rys. 2 a oryginał przedstawiono na Rys. 1. Porównanie oryginału z predykcją ukazuje, że ten model potrafił dobrze przewidywać te nuty, które były w powtarzających się sekwencjach, natomiast miał problem z właściwym wyznaczeniem długości trwania nut oraz dodawał nieistniejące nuty lub nie znajdywał istniejących, w miejscach gdzie było dużo dźwięków w krótkim momencie czasu.

Na koniec zbinaryzowane macierze predykcji tych modeli przekonwertowano na pliki MIDI. Niestety ten proces okazał się być trudniejszy niż przypuszczano i napotkano wiele problemów podczas ich tworzenia, np. zapisywał jedynie w postaci monofonicznej (nie było dwóch dźwięków w tym samym czasie).



Rys. 1. Utwór muzyczny w widoku piano roll



Rys. 2. Predykcja utworu muzycznego w widoku piano roll (model M1)

LITERATURA

[1] K. S. Michael Bereket, "An ai approach to automatic natural music transcription", 2017.
[2] Dixon, Dixon, Duan, Ewert,"Automatic Music Transcription – an overview", 2019
[3] Gu, Zhang, Xue, Wu, "Multi-scale Sub-band Constant-Q Transform Discriminator for High-filedlity Vocoder", 2023
[4] Emiya, Badeau, David, "Multipitch Estimation of Piano Sounds Using a New Probabilistic Spectral Smoothness Principle", 2010
[5] Librosa – audio and music processing in Python, (https://librosa.org)
[6] Pretty-midi documentation, (https://craffel.github.io/pretty-midi/)

Paper ID: **16**

BADANIE PORÓWNAWCZE GŁĘBOKICH MODELI AUTOMATYCZNEGO ROZPOZNAWANIA MOWY END-TO-END DLA ROZMÓW LEKARZ-PACJENT W JĘZYKU POLSKIM W RZECZYWISTYM ŚRODOWISKU AKUSTYCZNYM

Karolina Pondel-Sycz, Piotr Bilski
Instytut Radioelektroniki i Technik Multimedialnych, Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska
Autor korespondencyjny: **Karolina Pondel-Sycz**, *karolina.pondel.dokt@pw.edu.pl*

Słowa kluczowe: automatyczne rozpoznawanie mowy, głębokie modele neuronowe, transformer, conformer

1. WSTĘP

Celem badania było znalezienie najbardziej efektywnych metod automatycznego rozpoznawania mowy (ARM), które można wykorzystać do zbudowania systemu transkrypcji rozmowy lekarz-pacjent w języku polskim. W takim scenariuszu modele ARM napotykają wyzwania, takie jak dźwięki otoczenia, hałas, pogłos i nakładające się rozmowy, zawartość terminologii medycznej. W badaniu oceniono cztery modele ARM oparte o głębokie sieci neuronowe: XLSR-53 large Polish [1] , STT Pl Quartznet15x5 [3], STT Pl FastConformer Hybrid Transducer-CTC Large P&C [2] i Whisper-large [5]

2. ZASTOSOWANE METODY

W badaniu zastosowano autorski zbiór danych nagrań rozmów lekarz-pacjent i transkrypcji, o której wiadomo, że nie został wykorzystana do treningu żadnego z testowanych modeli, składający się z nagrań wywiadów medycznych, przeprowadzonych w warunkach rzeczywistych.

Nagrania wykonano za pomocą 5 mikrofonów, różnej jakości. Długość nagrań waha się od 1,36 do 6,34 minut (średnia 3,06 minuty). Pierwotnie nagrania wykonano jako 2-kanałowe, ze względu na wymagania modeli FastConformer i QuartzNet, nagrania zostały przekonwertowane na 1 kanałowe, z częstotliwością próbkowania 16 kHz.

Do oceny badanych modeli zastosowano dwie klasyczne metryki ARM – Word Error Rate (WER) i Character Error Rate (CER) oraz dodatkowe metryki, w tym: Match Error Rate (MER), Word Accuracy (WAcc), Word Information Preserved (WIP), Word Information Lost (WIL), odległość Levenshtein'a (Lev. Dist.), współczynnik Levenshtein'a (LR), podobieństwo Jaro – Winkler (JW sim.) oraz wskaźnik Jaccard'a [4], [6].

3. EKSPERYMENT

W eksperymencie wszystkie nagrania ze zbioru danych, podawano na wejście każdego modelu. Wynikiem była hipoteza w formie tekstowej. Tekst referencyjny i hipoteza zostały znormalizowane: przekształcono znaki na wielkie i usunięto wszystkie znaki specjalne. Utworzone w ten sposób teksty referencyjne i hipotezy zostały wykorzystane do obliczenia wskaźników wymienionych w sekcji 2.

Obliczono średnie wartości metryk wraz z odchyleniem standardowym, wygenerowano wykresy pudełkowe dla każdej metryki i przeprowadzono test ANOVA wpływu jakości mikrofonu na wynik rozpoznawania (hipotezę).

Metrics	FastConformer	Quartznet
WER %	46.30 (±12.33)	76.25 (±6.88)
CER %	33.18 (±12.29)	34.89 (±7.47)
MER %	45.66 (±12.36)	75.88 (±7.00)
WAcc %	53.70 (±12.33)	23.75 (±6.88)
WIP %	43.67 (±12.50)	8.77 (±4.03)
WIL %	56.33 (±12.50)	91.23 (±4.03)
Lev. dist.	808.96 (±559.80)	830.10 (±433.27)
LR	0.79 (±0.09)	0.74 (±0.06)
JW sim.	0.77 (±0.06)	0.82 (±0.04)
Jaccard	0.72 (±0.13)	0.83 (±0.04)
Metrics	**Wav2Vec**	**Whisper**
WER %	67.96 (±12.54)	20.84 (±9.53)
CER %	29.87 (±9.55)	13.89 (±6.59)
MER %	76.08 (±12.91)	19.68 (±8.23)
WAcc %	32.04 (±12.54)	79.16 (±9.53)
WIP %	15.74 (±9.81)	73.99 (±10.96)
WIL %	84.26 (±9.81)	26.01 (±10.96)
Lev. dist.	730.68 (±463.49)	356.40 (±271.86)
LR	0.79 (±0.08)	0.92 (±0.04)
JW sim.	0.83 (±0.05)	0.82 (±0.01)
Jaccard	0.85 (±0.06)	0.93 (±0.03)

Tab. 1. Średnie wartości metryk (z odchyleniem standardowym) dla testowanych modeli.

3. WNIOSKI

Model Whisper-large wypadł najlepiej dla wszystkich wskaźników, z wyjątkiem podobieństwa Jaro – Winkler, dla którego wynik jest zbliżony do Wav2Vec. Wyniki Whisper są mniej rozproszone, więc model popełnia błędy na mniejszą skalę. Analizując wyniki LR dla Whisper, można zauważyć, że w większości przypadków ciągi referencyjne i hipotetyczne są blisko siebie. Przewagę modelu Whisper podkreślają również uzyskane odległości Levenshtein'a, które są niższe niż w przypadku innych modeli. Wskazuje to, że do uzyskania tego samego tekstu, co w referencji, wymagana jest mniejsza liczba zmian w hipotezie. Minimalna wartość WER dla Whisper wyniosła 6,97%, co wskazuje, że model jest w stanie osiągnąć skuteczność zbliżoną do podawanej przez jego twórców.

LITERATURA

[1] CONNEAU, A., et al. *Unsupervised cross-lingual representation learning for speech recognition.* arXiv (Cornell University), Jun. 2020, doi: 10.48550/arXiv.2006.13979

[2] GULATI, A., et al. *Conformer: Convolution-augmented transformer for speech recognition,* arXiv (Cornell University), May 2020, doi: 10.48550/arXiv.2005.08100

[3] KRIMAN, S., et al. *Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions.* ICASSP 2020, pp. 6124-6128, Barcelona, Spain

[4] MORRIS, A.C., et al., *From wer and ril to mer and wil: improved evaluation measures for connected speech recognition*. In Proc. INTERSPEECH 2004, pages 2765–2768. ISCA Speech.

[5] RADFORD, A., et al. *Robust speech recognition via large-scale weak supervision.* International Conference on Machine Learning. PMLR, 2023, pp. 28492-28518, Seattle, WA, USA

[6] ROZINEK, O., MARES J., *Fast and precise convolutional jaro and jaro-winkler similarity*. In 35th Conf. of Open Innovations Association (FRUCT), pages 604–613, Tampere, Finland. IEEE.

Paper ID: **17**

ANALIZA INTONACJI CHÓRZYSTÓW W RÓŻNYCH WARUNKACH ODSŁUCHU: GŁOŚNIK, SŁUCHAWKI OTWARTE, SŁUCHAWKI ZAMKNIĘTE

Pietrzak Agnieszka Paula, Krawczyk Aleksandra
Instytut Radioelektroniki i Technik Multimedialnych, Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska
Autor korespondencyjny: **Agnieszka Paula Pietrzak**, *agnieszka.pietrzak@pw.edu.pl*

Poprawna intonacja muzyków śpiewających w chórze jest podstawą jakości brzmienia zespołu. W niniejszej pracy przeprowadzono analizę intonacji chórzystów w różnych warunkach odsłuchowych, które mogą mieć zastosowanie w przypadku studyjnych nagrań chóru: odsłuch głośnikowy, odsłuch przez słuchawki zamknięte oraz odsłuch przez słuchawki otwarte. Przeprowadzone badanie obejmowało analizę zaśpiewanych przez chórzystów dźwięków kwinty czystej do usłyszanego dźwięku referencyjnego oraz dźwięków tercji wielkiej uzupełniającej względem dźwięków referencyjnych tworzących kwintę czystą. W ramach pracy przeprowadzono analizę czasowo-częstotliwościową śpiewu chórzystów, w szczególności analizę częstotliwości podstawowej i błędów intonacji. Najwyższe wartości błędu intonacji zaobserwowano dla odsłuchu przez głośnik (20 centów), najniższe wartości wystąpiły dla słuchawek otwartych (17 centów), a w przypadku odsłuchu przez słuchawki zamknięte błąd wynosił 18 centów.

Słowa kluczowe: intonacja chórzystów, częstotliwość podstawowa, błędy intonacji, odsłuch słuchawkowy

Paper ID: **18**

ASSESSING MODELS FOR ESTIMATION ENSEMBLE WIDTH IN BINAURAL MUSIC RECORDINGS: ROBUSTNESS TO REVERBERATION AND NOISE

Paweł Antoniuk, Sławomir Krzysztof Zieliński
Faculty of Computer Science, Białystok University of Technology
Corresponding author: **Paweł Antoniuk**, *pawel.antoniuk@sd.pb.edu.pl*

Keywords: binaural audio, ensemble width, audio perception, localization, reverberation, machine learning

1. INTRODUCTION AND METHODOLOGY

Recent advances in software and consumer electronics have driven widespread adoption of binaural audio technology [2], leading to an anticipated increase in demand for spatial analysis tools. While existing music information retrieval algorithms are capable of effectively addressing the temporal and spectral characteristics of spatial audio recordings [4], there is a notable absence of methods capable of analyzing the spatial features of real-world binaural recordings. The development of such methods could potentially lead to the creation of spatial sound analysis tools, which could be utilized to retrieve spatial information from real-world binaural recordings.

Most studies on sound source localization in binaural recordings focus on the localization of individual sound sources in isolation [3]. While offering more information, the existing methods that exemplify this approach require a priori knowledge of the number of sources; moreover, these methods can analyze a relatively small number of concurrent sound sources. These limitations render these methods impractical in real-life scenarios, where such knowledge is not provided. In a series of recent studies, we employed an alternative approach, wherein sound sources were modeled as ensembles that could be described in terms of its width [1].

This work compares the following recent models for ensemble width prediction in terms of Mean Absolute Error (MAE): (1) a method based on an auditory model and decision trees, (2) a method using neural networks, and (3) a method leveraging spatial spectrograms [1]. For this purpose, the original models were tested on signals with predefined signal-to-noise ratios and with simulated rooms exhibiting different reverberation characteristics.

2. RESULTS

Under anechoic conditions without any interfering noise, experimental results revealed that the auditory system-based model (1) performed best with an MAE of 6.63° (±0.12°), followed by the neural network-based model (2) at 8.57° (±0.19°), and the spatial spectrogram model (3) at 13.62° (±0.93°). The differences between results are significant, with $p < 0.01$ for all comparisons. As shown in Figure 2, all models demonstrated limited resilience to noise, with model (2) exhibiting the highest robustness for $\mathrm{SNR} > -3$ dB, model (1) with $\mathrm{SNR} > 10$ dB, while model (3) showed lowest resilience for $\mathrm{SNR} > 60$ dB. The reverberation experiment indicated that none of the models exhibited significant robustness, with even the best-performing model (1) showing significant degradation at $\mathrm{RT60} = 0.1$ sec. This lack

of robustness can be attributed to the models not being trained on reverberant signals, suggesting a potential area for future improvement. Such enhancement can potentially lead to the development of objective assessment tools for audio engineers working with real-life binaural audio.

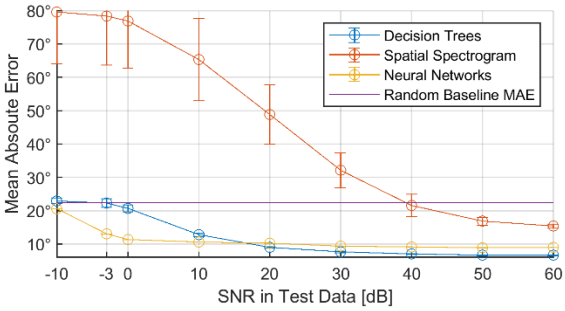


Fig. 2. Robustness of the models to noise.

REFERENCES

[1] ANTONIUK P, ZIELIŃSKI S. K., Blind estimation of ensemble width in binaural music recordings using ‘spatiograms’ under simulated anechoic conditions, Audio Engineering Society Conference, UK, 2023.
[2] BEGAULT D. R., 3-D Sound for Virtual Reality and Multimedia, NASA Center for AeroSpace Information: Hanover, MD, USA, 2000.
[3] MAY T., MA N., BROWN G. J., Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Australia, 2015, 2679–2683.
[4] SIMONETTA F., NTALAMPIRAS S., AVANZINI F., Multimodal Music Information Processing and Retrieval: Survey and Future Challenges in Proceedings of the International Workshop on Multilayer Music Representation and Processing (MMRP), Milan, Italy, 2019, 10-18.

Paper ID: **19**

DŹWIĘK 3D W APLIKACJACH INTERNETOWYCH

Marcin Lewandowski, Jan Radzimiński
Instytut Radioelektroniki i Technik Multimedialnych, Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska
Autor korespondencyjny: **Marcin Lewandowski**, *marcin.lewandowski@pw.edu.pl*

Wirtualna i rozszerzona rzeczywistość wykorzystują technologię dźwięku 3D w tworzeniu trójwymiarowych doświadczeń, a aplikacje internetowe coraz częściej oferują możliwość odtwarzania dźwięku przestrzennego w czasie rzeczywistym. Głównym celem pracy jest przeprowadzenie dokładnej analizy i porównania wybranych technologii służących do symulowania dźwięku 3D w odsłuchu słuchawkowym, przeznaczonych do implementacji właśnie w aplikacjach internetowych. Analiza jest podzielona na część testów wydajnościowych przeprowadzonych w różnych scenariuszach działania aplikacji oraz oceny cech wrażeniowych dźwięku wśród grupy słuchaczy. Uczestnicy testów słuchowych zostali poproszeni o opisanie swoich doświadczeń i identyfikację kierunków dochodzenia dźwięków pochodzących z różnych lokalizacji w przestrzeni za pomocą specjalnie zaprojektowanej platformy testowej. Porównywane były biblioteki WebAudio API, Resonance Audio, JS Ambisonics i Mach1 Audio SDK. Biblioteka Mach1 Audio SDK okazała się najbardziej obciążająca procesor oraz pamięć, ale oferowała lepszą jakość dźwięku oraz dokładność odtwarzania kierunków dochodzenia dźwięku niż inne silniki przetwarzania.

Słowa kluczowe: dźwięk przestrzenny, aplikacje internetowe

Paper ID: **20**

NIERÓWNOMIERNE PRÓBKOWANIE PRZESTRZENNE W ZAGADNIENIU ESTYMACJI KIERUNKU NADEJŚCIA (DOA) FALI AKUSTYCZNEJ

Zbigniew Świętach, Bogusław Szlachetko, Przemysław Plaskota, Bartłomiej Kruk, Michał Łuczyński, Jędrzej Szczepaniak
Katedra Akustyki, Mutlimediów i Przetwarzania Sygnałów, Wydział Elektroniki, Fotoniki i Mikrosystemów, Politechnika Wrocławska
Autor korespondencyjny: **Zbigniew Świętach**, *zbigniew.swietach@pwr.edu.pl*

Słowa kluczowe: kierunek nadejścia fali (DOA), formowanie wiązki, macierz mikrofonów

WPROWADZENIE

W niniejszej pracy zaprezentowana zostanie metoda wyznaczania kierunku nadejścia fali akustycznej (DOA), gdzie kierunek rozumiany jest jako azymut i elewacja poruszającego się obiektu. Za pomocą macierzy mikrofonów, wzmacniaczy, przetworników ADC (przetworników analogowo-cyfowych) oraz autorskiego oprogramowania wyznacza się przestrzenną „mapę" chwilowego położenia obiektów latających w przestrzeni. Wszystko to realizowane jest za pomocą typowych urządzeń stosowanych w technice audio, tzn. mikrofonów firmy Superlux oraz 16 kanałowego procesora dźwięku firmy Roland. Dane w formie strumieni PCM (modulacja kodowo-impulsowa) zapisywane są w plikach *.wav i przesyłane do dalszej obróbki w komputerze, za pomocą środowiska obliczeniowego Matlab. Wybór Matlaba do dalszego przetwarzania sygnałów podyktowany jest prostotą i przejrzystością kodu źródłowego, będącego plikiem tekstowym. Ponadto sposób zapisu problemu technicznego czy matematycznego w Matlabie jest wręcz intuicyjny oraz niewiele odbiega od zapisu takiego problemu przy użyciu standardowej notacji matematycznej.

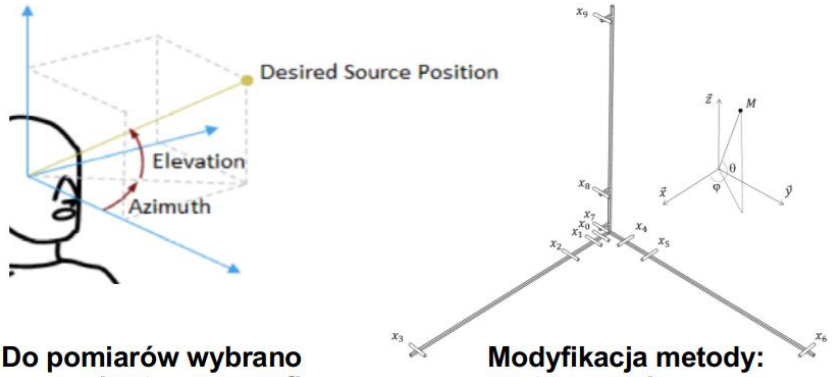
Obiektami latającymi, będącymi przedmiotem testów są dwa drony oraz dodatkowo samoloty pasażerskie, przelatujące w odległości około 2-3 kilometrów od macierzy mikrofonów. Wykonano szereg pomiarów fizycznych, zarówno w komorze akustycznej (symulacja fali akustycznej generowanej przez dron na pomocą głośników) jak i w otwartym terenie przy użyciu latającego drona.

Zastosowanie technik audio do wyznaczania DOA [1] jest uzupełnieniem technik radarowych wykrywania obiektów latających, przy czym akustyczne DOA umożliwia pasywne wykrywanie np. niewielkich dronów znajdujących się na niskim pułapie i w niewielkiej odległości od potencjalnego celu [3]. W opisanej sytuacji metody radarowe są mało skuteczne.

Ważnym celem omawianego eksperymentu było również zbadanie możliwości wykorzystania macierzy mikrofonów o innej niż jednorodna strukturze przestrzennej [2]. To oznacza nierównomierne próbkowanie przestrzenne i nie było oczywiste, jak takie rozmieszczenie mikrofonów wpłynie na możliwość estymacji DOA. Finalnie okazało się, że stosując metodę estymacji DOA opartą na formowaniu wiązki w dziedzinie czasu (DSB – delay and sum beamforming), ograniczenia dotyczące próbkowania przestrzennego nie znalazły tutaj zastosowania. Było to zgodne z przeprowadzoną przed pomiarami analizą teoretyczną. Wzmiankowane ograniczenia są jednak nadal istotne jeżeli do estymacji DOA używa się metod widmowych [4].

REALIZACJA AKUSTYCZNEGO DOA

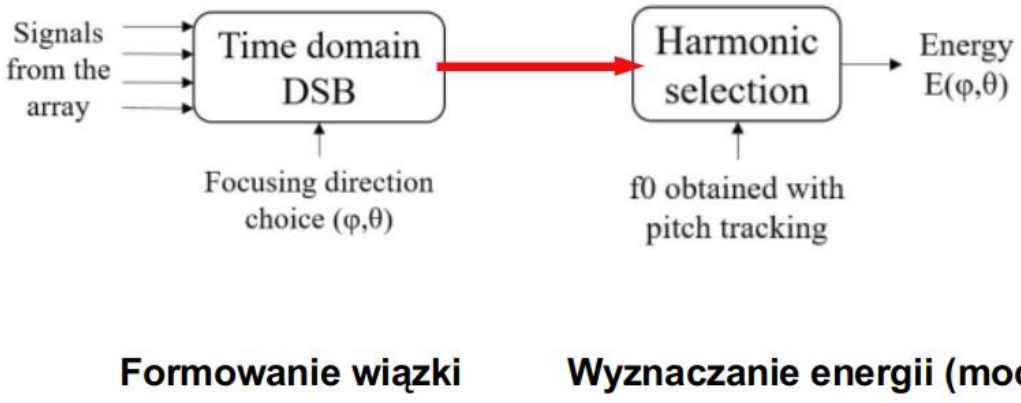
Na ryc. 1 pokazano schemat współrzędnych przestrzennych i rozmieszczenie mikrofonów w macierzy pomiarowej.



Ryc. 1 Układ współrzędnych pomiarowych i macierz mikrofonów

Pomiary przeprowadzono przy użyciu 10 mikrofonów rozmieszczonych nierównomiernie w macierzy pomiarowej. Mikrofony numerowane są od zera do dziewięciu, przy czym numer zero oznacza mikrofon referencyjny.

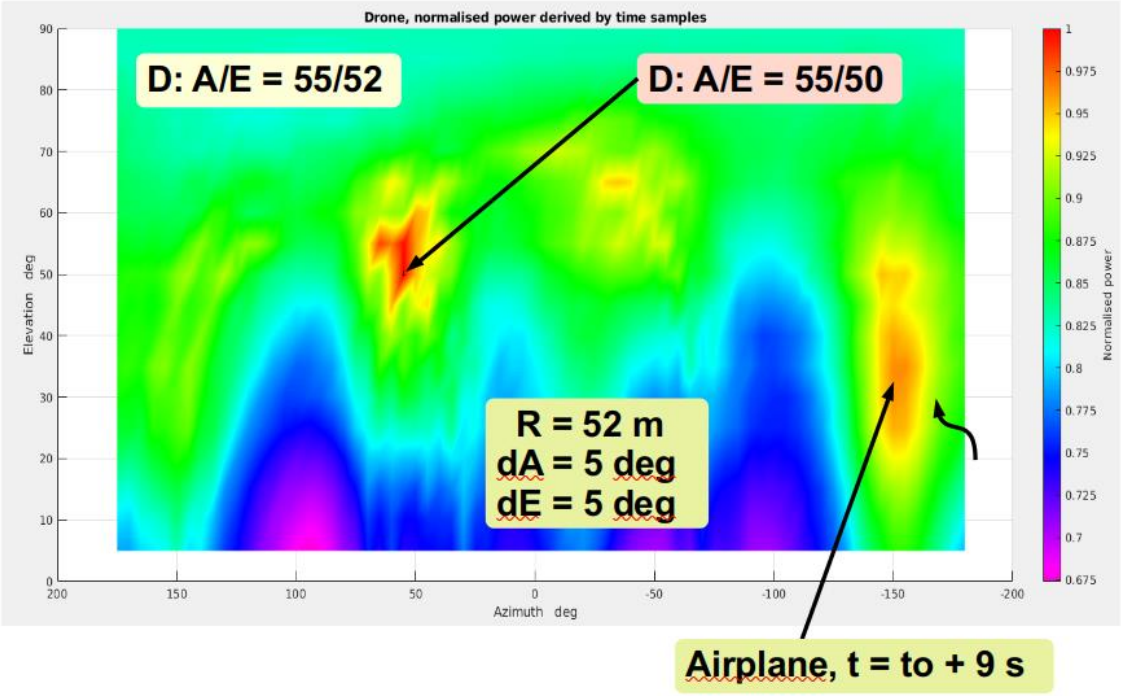
Na ryc. 2 zamieszczono schemat blokowy realizacji akustycznego DOA. Pierwszy blok odpowiada za formowanie wiązki na podstawie sygnałów otrzymanych z mikrofonów. Wiązka formowana jest za pomocą typowej metody czasowego beamformingu (DSB) [1]. Takie podejście umożliwia stosowanie dowolnej konfiguracji przestrzennej mikrofonów pomiarowych. Nie występuje tutaj ograniczenie związane z próbkowaniem przestrzennym.



Ryc. 2 Schemat blokowy: blok beamformera i blok wyznaczania energii uformowanej wiązki

Ryc. 3 przedstawia jeden z wyników eksperymentów przeprowadzonych z wykorzystaniem mobilnych źródeł dźwięku (BSP). Eksperymenty przeprowadzono w środowisku naturalnym. Poza BSP użytym w eksperymentach pojawiły się zakłócenia pochodzące od przelatujących samolotów pasażerskich.

W takich przypadkach macierz umożliwiała określenie DOA obydwu obiektów jednocześnie.



Ryc. 3 Mapa przestrzenna rozkładu mocy uformowanych wiązek. Dron znajduje się w położeniu A/E = 55/50 deg, natomiast przelatujący samolot w położeniu A/E = -155/35 deg.

Po przeprowadzeniu dotychczasowych prac można wyciągnąć następujące wnioski:

1) Akustyczne DOA działa poprawnie i umożliwia wykrywanie kilku obiektów jednocześnie

2) System nie wymaga równomiernego rozmieszczenia mikrofonów, zatem równomierne próbkowanie przestrzenne nie jest konieczne dla czasowych metod formowania wiązki

LITERATURA

[1] Nathan Itare, Jean-Hugh Thomas, Kosai Raoof, Torea Blanchard Acoustic Estimation of the Direction of Arrival of an Unmanned Aerial Vehicle Based on Frequency Tracking in the Time-Frequency Plane. MDPI Sensors 26.05.2022

[2]Sedunov, A.; Haddad, D.; Salloum, H.; Sutin, A.; Sedunov, N.; Yakubovskiy, A. Stevens Drone Detection Acoustic System and Experiments in Acoustics UAV Tracking. In Proceedings of the 2019 IEEE International Symposium on Technologies for Homeland Security (HST), Woburn, MA, USA, 5–6 November 2019; pp. 1–7.

[3] Kloet, N.; Watkins, S.; Clothier, R. Acoustic signature measurement of small multi-rotor unmanned aircraft systems. Int. J. Micro Air Veh. 2017, 9, 3–14.

[4] Zahorian, S.A.; Hu, H. A spectral/temporal method for robust fundamental frequency tracking. J. Acoust. Soc. Am. 2008, 123, 4559–4571.

Paper ID: **21**

RELACJA CZASU WCZESNEGO ZANIKU EDT DO CZASU POGŁOSU RT W ZALEŻNOŚCI OD RODZAJU DEKORACJI STOSOWANEJ W TEATRZE DRAMATYCZNYM

Piotr Z. Kozłowski
Katedra Akustyki, Mutlimediów i Przetwarzania Sygnałów, Wydział Elektroniki, Fotoniki i Mikrosystemów, Politechnika Wrocławska
Autor korespondencyjny: **Piotr Kozłowski**, *piotr.kozlowski@pwr.edu.pl*

Scenografia wykorzystywana podczas spektakli dramatycznych ma istotny wpływ na warunki pogłosowe panujące na scenie oraz widowni. Dla postrzegania pogłosowości sali istotna jest nie tylko wartość samego czasu pogłosu ale także jego relacja do czasu wczesnego zaniku. W pracy przedstawiono wyniki badań przeprowadzonych w Teatrze Polskim we Wrocławiu dla kilku odmiennych rodzajów scenografii. Na podstawie analizy zebranych wyników zaprezentowano konkluzje dotyczące wpływu dekoracji na pogłosowość sali oraz relację EDT / RT.

Słowa kluczowe: teatr dramatyczny, pogłos, dekoracja

Paper ID: **22**

LOKALIZACJA ŹRÓDŁA DŹWIĘKU DLA RÓŻNYCH RENDERERÓW BINAURALNYCH

Magdalena Piotrowska(*), Olga Krzyżyńska, Paweł Małecki
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
Autor korespondencyjny: **Magdalena Piotrowska**, *mpiotrowska@agh.edu.pl*

Artykuł porusza zagadnienie lokalizacji pozornego źródła dźwięku dla różnych rendererów binauralnych, w kontekście przetwarzania formatów dźwięku przestrzennego na potrzeby odsłuchu słuchawkowego. W ramach badań przeanalizowano precyzję lokalizacji źródła dźwięku dla technologii ambisonicznej oraz dwóch różnych wersji renderera z formatu Dolby Atmos. Próbki dźwiękowe odtwarzano uczestnikom w środowisku wirtualnej rzeczywistości (VR), co umożliwiło naturalne wskazywanie kierunków dźwięku. Wyniki wskazują, że renderer Dolby Atmos charakteryzuje się największą dokładnością w odwzorowaniu położenia źródła dźwięku, szczególnie w przypadku próbek pochodzących zza głowy słuchacza. Badania pokazały również, że słuchacze lepiej rozpoznają położenie dźwięku w płaszczyźnie horyzontalnej (lewo-prawo) niż wertykalnej (góra-dół). Zastosowanie VR w badaniach psychoakustycznych okazało się nie tylko skuteczne, ale również interesujące dla uczestników, co podkreśla potencjał tej technologii do dalszych badań nad dźwiękiem przestrzennym.

Słowa kluczowe: Lokalizacja dźwięku Odsłuch binauralny Dźwięk przestrzenny Renderery binauralne Wirtualna rzeczywistość (VR)

Paper ID: **23**

TECHNICZNE PARAMETRY A PERCEPCJA: SUBIEKTYWNE I OBIEKTYWNE PODEJŚCIE DO OCENY WPŁYWU PRZEWODÓW GŁOŚNIKOWYCH NA DŹWIĘK

Tomasz Kopciński, Dominika Kuczak, Bartłomiej Kruk, Tomasz Nowak
Katedra Akustyki, Mutlimediów i Przetwarzania Sygnałów, Wydział Elektroniki, Fotoniki i Mikrosystemów, Politechnika Wrocławska
Autor korespondencyjny: **Tomasz Kopciński**, *tomasz.kopcinski@pwr.edu.pl*

Przewód głośnikowy służący do połączenia wzmacniacza mocy z urządzeniem głośnikowym jest tematem wielu sporów pomiędzy badaczami, a grupą miłośników muzyki zwanych audiofilami. W celu analizy tego zagadnienia przeprowadzono pomiary obiektywne oraz subiektywne, w których badano wpływ zmiany rodzaju przewodu głośnikowego na wrażenia słuchowe odbiorcy. Wyniki badań obiektywnych obejmują charakterystyki częstotliwościowe badanych przewodów, wartości THD+N oraz pomiary charakterystyk częstotliwościowych zestawu głośnikowego podłączonego do wzmacniacza za pomocą różnych przewodów głośnikowych. Do badań subiektywnych wykorzystano 15-sekundowe próbki muzyczne z trzech gatunków muzycznych oraz komparatywną metodę oceny. Badanymi cechami wrażeniowymi były: dynamika, jasność brzmienia, przejrzystość oraz ocenę ogólną.

W testach obiektywnych nie zauważono różnic występujących pomiędzy badanymi przewodami. Można więc stwierdzić, iż przewody głośnikowe wybierane do podłączania domowych zestawów głośnikowych nie mają znaczenia pod względem zmiany brzmienia systemu stereofonicznego. Zbieżne wnioski można wyciągnąć z pomiarów subiektywnych. Test Levene'a wykazał, iż oceny były homogeniczne, z kolei analiza ANOVA nie wykazała istotności statystycznej. W wynikach dla gatunków muzycznych zaobserwowano jedynie nieznaczne różnice.

Słowa kluczowe: przewód głośnikowy, pomiary, badania subiektywne, ANOVA

Paper ID: **24**

WPŁYW WYGRZEWANIA PRZETWORNIKÓW NA PARAMETRY SUBIEKTYWNE I OBIEKTYWNE ZESTAWÓW GŁOŚNIKOWYCH.

Paweł Kuflowski, Tomasz Kopciński, Dominika Kuczak, Tomasz Nowak
Katedra Akustyki, Mutlimediów i Przetwarzania Sygnałów, Wydział Elektroniki, Fotoniki i Mikrosystemów, Politechnika Wrocławska
Autor korespondencyjny: **Paweł Kuflowski**, *259607@student.pwr.edu.pl*

Wygrzewanie głośników to kontrowersyjny temat w świecie audio, opierający się na założeniu, że nowe urządzenia muszą osiągnąć stabilne parametry po okresie użytkowania. Zwolennicy twierdzą, że wygrzewanie poprawia elastyczność elementów, co pozytywnie wpływa na brzmienie, zwłaszcza w zakresie małych częstotliwości, redukuje początkowe zniekształcenia oraz poprawia charakterystykę częstotliwościową głośnika i precyzję w odwzorowaniu dynamiki. Jednak przeciwnicy wskazują na brak naukowych dowodów na słyszalne zmiany a dodatkowo zarzucają możliwość powstawania efektu placebo podczas badań subiektywnych. Ich zdaniem, nowoczesna technologia produkcji i możliwość precyzyjnej kontroli jakości wyklucza możliwość powstania słyszalnych zmian w trakcie użytkowania urządzenia. Poruszony został aspekt procesów starzeniowych elementów głośnika, aby lepiej zrozumieć zachodzące zmiany

w materiałach, z których został stworzony. Przytoczona polska norma opisuje taki zabieg jak wygrzewanie wstępne i określa dokładne warunki i czas, w jakich cały proces powinien zostać przeprowadzony. Przytoczone w pracy publikacje skłaniają do twierdzenia, że wygrzewanie przetworników, faktycznie prowadzi do drobnych zmian w elastyczności materiałów, jednak są one niesłyszalne dla przeciętnego słuchacza.

Przeprowadzone w pracy badania obiektywne wykazały, że zachodzi nieznaczne przesunięcie maksimum impedancji głośnika oraz zmiany w odpowiedzi amplitudowej. Widoczne są również duże różnice pomiędzy zestawami tej samej serii, co wynika z warunków na liniach produkcyjnych i utrzymywanych dopuszczalnych rozbieżności. Natomiast badania subiektywne ukazują, że w przypadku wszystkich testów porównawczych, urządzenia głośnikowe są oceniane jednorodnie.

Słowa kluczowe: wygrzewanie przetworników

Paper ID: **25**

ODPORNA PARAMETRYZACJA MOWY OPARTA NA SYNCHRONIZACJI METOD CEPSTRALNYCH Z OKRESEM PODSTAWOWYM TONU KRTANIOWEGO.

Stanislaw Gmyrek, Robert Hossa
Katedra Akustyki, Mutlimediów i Przetwarzania Sygnałów, Wydział Elektroniki, Fotoniki i Mikrosystemów, Politechnika Wrocławska
Autor korespondencyjny: **Stanisław Gmyrek**, *stanislaw.gmyrek@pwr.edu.pl*

Słowa kluczowe: odporna parametryzacja cepstralna, okres podstawowy, korekcja widma amplitudowego

1. WPROWADZENIE

W systemach przetwarzania mowy (SPS) istnieje potrzeba kompensacji negatywnego wpływu wielu czynników, takich jak techniczne warunki rejestracji, zmienność wewnątrz- i międzyosobnicza, kontekstowość itp., które negatywnie wpływają na wydajność systemów. Niniejsza praca rozważa problem odpornej parametryzacji. Spośród co najmniej kilkunastu różnych metod parametryzacji dostępnych w literaturze, do najczęściej wykorzystywanych i skutecznych w praktycznych zastosowaniach należą metody wykorzystujące transformacje czasowo-częstotliwościowe oraz reprezentacje cepstralne. Do tej grupy rozwiązań możemy zaliczyć algorytmy MFCC (Mel Frequency Cepstral Coefficients), HFCC (Human Factor Cepstral Coefficients), BFCC (Basiliar-membrane Frequency-band Cepstral Coefficient), GTCC (Gammatone Cepstral Coefficient) oraz AMS (Amplitude Modulation Spectrum). Istnieją także inne grupy rozwiązań wykorzystujących metody predykcji liniowej, a przykładami ich implementacji są parametryzacje LPCC (Linear Prediction Cepstral Coefficients) i PLP (Perceptual Linear Prediction). Większość z wyżej wymienionych metod parametryzacji posiada naturalnie mechanizmy zapewniające odporność na niewielkie zakłócenia szumowe. W celu poprawy skuteczności istniejących metod parametryzacji zaproponowano uzupełnienie ich o algorytm RASTA (Relative Spectra), który usuwa te składowe, które nie są związane z artykulacją mowy (RASTA-PLP).

1.2. TEORIA I CEL PRACY

Ogólnie rzecz biorąc, dźwięczne fragmenty sygnału mowy są kształtowane przez sygnał pobudzenia impulsu krtaniowego i trakt głosowy. Niestety, quasi-okresowość pobudzenia tonu krtaniowego, w przypadku parametryzacji opartych na reprezentacjach czasowo-częstotliwościowych, jest jednym z kluczowych czynników wpływających na znaczny rozrzut wartości wektora cech, wprowadzając zafalowania do widma amplitudowego (ang. ripples). W artykule zaproponowano rozwiązanie modyfikacji klasycznych metod cepstralnych o analizę ramek o zmiennej długości (zsynchronizowanej z okresem podstawowym T0) w celu znacznego zmniejszenia tego niepożądanego zjawiska. Pierwszym krokiem jest estymacja okresu podstawowego T0 w ramce sygnału mowy, a następnie wyznaczenie widma amplitudowego za pomocą STFT (Short Time Fourier Transform) ze zmiennym w czasie oknem o długości zgodnej z aktualną wartością T0.

Literatura pokazuje, że parametryzacja HFCC charakteryzuje się większą odpornością na zakłócenia

niż MFCC, a badania wykazały różnice w skuteczności rozpoznawania do 30 %. W rezultacie klasyczne rozwiązanie, tj. parametryzacja HFCC, zostało wybrane jako reprezentatywne dla naszego eksperymentalnego badania nad redukcją zafalowań w widmie amplitudowym i jego konsekwencjami. Celem takiej analizy było sprawdzenie statystycznych właściwości zestawu wektorów HFCC dla poszczególnych samogłosek w oparciu o wariancje ich składowych. Ponadto, w celu oceny skuteczności proponowanej metody, modele statystyczne w postaci GMM (Gaussian Mixture Model) dla poszczególnych fonemów w mowie polskiej zostały obliczone i przeanalizowane z wykorzystaniem koncepcji probabilistycznej miary odległości między nimi z korektą i bez korekty. Na koniec przeprowadzono ocenę skuteczności proponowanego podejścia poprzez porównanie miary błędu klasyfikacji Frame Error Rate (FER).

2. PODSUMOWANIE

Zaproponowana w niniejszej pracy modyfikacja parametryzacji HFCC za pomocą STFT zsynchronizowanej z okresem podstawowym spełnia założone oczekiwania. Poprzez estymację T0 i zmienną długość okna transformaty DFT zgodną z aktualną wartością tego okresu skutecznie eliminujemy wpływ quasi-okresowości sygnału pobudzenia na widmo amplitudowe. Co więcej, na podstawie wariancji współrzędnych wektorów cech można zaobserwować zmniejszenie obszaru zajmowanego przez te wektory dla każdej samogłoski oraz wzrost odległości Kullbacka-Leiblera pomiędzy rozkładami GMM samogłosek mowy polskiej. W konsekwencji eksperymenty z proponowaną metodą parametryzacji cepstralnej zsynchronizowanej z okresem podstawowym skutkują wzrostem efektywności systemu klasyfikacji samogłosek. Zmniejszenie wariancji estymatorów wektorów obserwacji i zmniejszenie błędów klasyfikacji zaobserwowano dla wszystkich analizowanych stanów. Dla klasyfikatora GMM poprawa ta sięga kilku procent. Obserwacje te jednak wyraźnie sugerują dalsze poszukiwania nowych skutecznych klasyfikatorów dla systemów technologii mowy. Jednocześnie należy pamiętać, że na zmienność składowych wektora cech, oprócz wpływu quasi-okresowości pobudzenia, wpływa szereg innych czynników, takich jak zmienność wewnątrz- i międzyosobnicza, kontekstowość, czy techniczne warunki rejestracji.