XXVIII Seminarium

ZASTOSOWANIE KOMPUTERÓW W NAUCE I TECHNICE 2018

Oddział Gdański PTETiS

doi: 10.32016/1.60.28

REJESTRACJA, PARAMETRYZACJA I KLASYFIKACJA ALOFONÓW Z WYKORZYSTANIEM BIMODALNOŚCI

Szymon ZAPOROWSKI[1], Sebastian CYGERT[1] , Grzegorz SZWOCH[1], Grażina KORVEL[2], Andrzej CZYŻEWSKI[1]

1. Politechnika Gdańska, Wydział Elektroniki, Telekomunikacji i Informatyki, Katedra Systemów Multimedialnych tel.: 58-348-6332, e-mail: smck@multimed.org
2. Institute of Data Science and Digital Technologies, Vilnius University, Lithuania e-mail: grazina.korvel@mii.vu.lt

Streszczenie: Praca dotyczy rejestracji i parametryzacji alofonów w języku angielskim z wykorzystaniem dwóch modalności. W badaniach dokonano rejestracji wypowiedzi w języku angielskim mówców, których znajomość tego języka odpowiada poziomowi rodowitego mówcy. W kolejnym etapie wyodrębnione zostały alofony z nagrań fonicznych i odpowiadające im sygnały wizyjne. W procesie tworzenia wektorów cech wykorzystano odrębne systemy parametryzacji, osobne dla każdej modalności. Do parametryzacji sygnału fonicznego użyto typowych deskryptorów stosowanych w obszarze rozpoznawania mowy i muzyki. W nagraniach z systemu przechwytywania ruchu zaproponowano własne rozwiązania. Do klasyfikacji alofonów wykorzystano sieci neuronowe oraz maszynę wektorów nośnych w podejściu jedno- i dwumodalnym. Stwierdzono, że skuteczność rozpoznawania wzrasta wraz z wykorzystaniem więcej niż jednej modalności.

Słowa kluczowe: Sieci neuronowe, klasyfikacja, facial motion capture.

1. WSTĘP

Obecnie obserwuje się znaczny wzrost popularności rozwiązań związanych z automatycznym rozpoznawaniem mowy (ASR – *Automatic Speech Recognition*). Główną metodą klasyfikacji i wykrywania stosowaną w tej dziedzinie jest uczenie maszynowe [1][2][3]. Większość tych rozwiązań jest znana od lat i szczegółowo opisana w licznych pracach [1][2][3]. Brakuje jednak prac nad próbą klasyfikacji elementów mowy na poziomie alofonicznym. Artykulacja alofonów jest złożonym procesem, dlatego klasyfikacja pojedynczych alofonów jest znacznie trudniejsza. Szczególnie zależność alofonu od poprzednich dźwięków, kontekstu wypowiedzi, krótkiego czasu trwania i indywidualnych cech wymagają zmiany podejścia znanego z literatury. W związkuz tym zdecydowano się na użycie technologii *Facial Motion Capture* (FMC), jako głównego nośnika informacji związanego z artykulacją dźwięków alofonicznych i przetestowanie przydatności takiego podejścia. Innym wykorzystanym rozwiązaniem jest połączenie parametrów audio, a mianowicie parametrów mel-cesptralnych (MFCCs - *Mel Frequency Cepstrum Coefficients*) i współczynników predykcji liniowej (LPC - *Linear Predictive Coding*).

W literaturze opisano wykrywanie ruchu warg za pomocą obrazu wideo [4]. Obecnie nie jest problematyczne określenie regionu zainteresowania [5] (ROI – *Region of Interest*) lub głębszego przetwarzania obrazu. Problem stanowi ilość danych ograniczona z przyczyn praktycznych, a mianowicie poprzez brak możliwości nagrania dużej liczby mówców oklejonych markerami w czasie zbliżonym do czasu nagrania korpusu audio-video. Algorytmy automatycznego rozpoznawania mowy, takie jak głębokie uczenie sieci neuronowych, są obecnie trenowane z wykorzystaniem danych zebranych z setek lub tysięcy mówców. Innym problemem jest ustawienie twarzy mówcy w położeniu prostopadłym do kamery. Każda zmiana kąta widzenia kamery powoduje przesunięcia, które w przypadku obrazu twarzy mówcy w 2D znacznie wpływają na uzyskane wyniki. Zastosowanie FMC ze znacznikami rozmieszczonymi w przestrzeni 3D na powierzchni twarzy pozwala na uzyskanie współrzędnych w trzech wymiarach. W ten sposób problem z perspektywą widzenia zostaje znacznie zredukowany. Wykorzystanie FMC powoduje jednak szereg trudności, w szczególności związanych z umieszczaniem znaczników na twarzach mówców i potrzebą jednolitego oświetlenia całej twarzy.

Celem autorów tej pracy jest zbadanie możliwości klasyfikacji alofonów z użyciem pojedynczych modalności oraz ich fuzji jako podstawy do dalszych prac nad systemami automatycznej transkrypcji fonetycznej mowy w języku angielskim.

2. DANE

2.1. Sesje nagraniowe

Nagrania wykonano w przystosowanym pomieszczeniu z wykorzystaniem ustrojów akustycznych w celu ograniczenia niepożądanych dźwięków. Do nagrania sygnału audio wykorzystano dwa mikrofony: krawatowy oraz o charakterystyce superkierunkowej, podłączone do zewnętrznego rejestratora. Sygnał został nagrany z częstotliwością próbkowania 48 kHz i z rozdzielczością 16 bitów. Do przechwycenia danych FMC użyto sześciu kamer Vicon Vero, podczas gdy wideo zostało nagrane przy użyciu jednej kamery referencyjnej Vicon (120fps) i aparatu cyfrowego (50 klatek na sekundę) oraz kamery sportowej (240 klatek na sekundę). Przed każdym użyciem system został skalibrowany, ponieważ pomiar fizycznej pozycji znaczników jest niezbędny do prawidłowego ustalenia pozycji markerów w programie.

Zastosowano 32 markery, z których 20 umieszczono na wargach mówców. Jednocześnie 4 spośród markerów

zostały umieszczone na specjalnej czapce, która była punktem odniesienia w stabilizowaniu obrazu FMC na etapie przetwarzania końcowego. Rozmieszczenie markerów przedstawiono na rysunku 1. Układ współrzędnych został zorientowany w taki sposób, że twarz mówcy została ustawiona równolegle do płaszczyzny XY, z współrzędnymi osi Y skierowanymi w stronę czoła i podbródka oraz z osią X biegnącą w kierunku uszu. Oś Z skierowana została w stronę kamer.

Rys. 1. Rozmieszczenie markerów na twarzy mówcy

Sesje nagraniowe podzielono na dwa dni, pozyskano dane od sześciu mówców. Każdy mówca wypowiedział 300 słów lub krótkich wyrażeń specjalnie wyselekcjonowanych przez współpracującego fonologa, aby zawierały odpowiednie warianty alofoniczne. W prezentowanym badaniu wykorzystano wyłącznie wybrane samogłoski i dyftongi. Wynika to z faktu, że posiadają one wyraziste obrazowanie w systemie FMC, dlatego po konsultacji ze specjalistą z dziedziny fonologii zdecydowano się nie wykorzystywać w analizach spółgłosek. Z zasad fonologii wynika, że samogłoski mają lepiej widoczne cechy artykulacyjne, które są łatwiej zauważalne przy użyciu nagrań FMC.

Podczas sesji nagraniowej napotkano kilka problemów. Pierwszym problemem, jak wspomniano wcześniej, było oświetlenie pomieszczenia. Światło w widmie widzialnym może powodować tworzenie artefaktów refleksyjnych, które wpływają na zapis FMC i które mogą, z kolei, tworzyć obrazy fałszywych znaczników. Dlatego przy użyciu sygnału podczerwieni, z której korzysta system FMC, nie można było efektywnie zarejestrować danych bez doświetlenia nagrywanej sceny. Co więcej, bez dodatkowego oświetlenia nie jest możliwe nagrywanie filmów wysokiej jakości, które były kluczowe z punktu widzenia specjalistów z zakresu fonetyki w wykrywaniu i etykietyzacji konkretnych samogłosek. Konieczne więc było użycie zarówno oświetlenia sufitowego, jak i reflektora skierowanego w stronę twarzy mówcy. Kolejnym problemem były znaczące różnice w kształcie ust mówców. Spowodowało to konieczność użycia nieco innego sposobu umieszczania znaczników dla każdej z nagrywanych osób.

2.2. Akwizycja danych

Przygotowane nagrania zostały wykorzystane przez specjalistów z dziedziny fonetyki do ręcznego oznaczania alofonów i dyftongów. Ze względu na brak możliwości automatyzacji procesu, etykietowanie zajęło około tygodnia przetwarzania dla każdego mówcy. Był to proces całkowicie manualny. Jest to jedna z największych wad takiego podejścia, ponieważ pozyskanie większej ilości danych jest procesem wyjątkowo czasochłonnym, nie pozwala w konsekwencji na zastosowanie algorytmu decyzyjnego opartego na głębokim uczeniu. Ponadto, różnice w czasie wypowiedzi wypowiadanych fragmentów były również problematyczne. Segmenty mowy mogą się znacznie różnić pod względem czasu trwania, np. mniej niż 40 ms dla samogłosek do prawie 400 ms dla dyftongów.

Nagrania wszystkich modalności zostały poddane edycji i podzielone na sekcje alofonów przez ekspertów z dziedziny fonologii, a następnie sparametryzowane przy użyciu zestawu deskryptorów. Liczność otrzymanych grup alofonów znacząco się od siebie różni,od 60 alofonów do tylko 1, co utrudnia zastosowanie algorytmów uczących się.

3. PARAMETRYZACJA

3.1. Normalizacja przesunięć

Pierwszą próbą sparametryzowania sygnału FMC było stworzenie prostego, dwuparametrowego podejścia. Jako parametry wykorzystano szerokość i wysokość otwartych ust mówcy. Taki sposób parametryzacji występuje w literaturze, jest jednym z podstawowych podejść w przypadku parametryzacji ust na podstawie obrazu video [5].

Wyniki klasyfikacji z użyciem tego rodzaju parametryzacji przy użyciu opisanej w pracy sieci neuronowej przedstawiono w kolejnym rozdziale w tablicy numer 1.

Ze względu na niewystarczającą dokładność tego rodzaju parametryzacji opracowano inną metodę. Została ona dokładnie opisana w innej pracy autorów dotyczącej tematyki klasyfikacji głosek alofonicznych [6]. Jest to parametryzacja bazująca na przesunięciach z wykorzystaniem współrzędnych kartezjańskich w celu zmiany pozycji markerów. Ten rodzaj parametryzacji został wykorzystany w przykładach i w wynikach pokazanych poniżej.

3.2. Parametryzacja sygnału fonicznego

W parametryzacji sygnału mowy stosuje się zróżnicowane parametry. Na podstawie eksperymentów przeprowadzonych w ramach badania alofonów [7][8] stworzono zestaw cech składających się z deskryptorów akustycznych, ich pochodnych i wartości statystycznych, które niosą informację. Taki wektor parametrów obejmuje parametry zarówno z dziedziny czasu i częstotliwości. Parametry w dziedzinie czasu obejmują: środek ciężkości (*Temporal Centroid* - TC), liczbę przejść przez zero (*Zero Crossing* - ZC), energię RMS (*Root Mean Square Energy*), czy wartość szytową do wartości RMS (*Peak to RMS*). Parametry związane z dziedziną częstotliwości uzyskano przez przekształcenie sygnału z dziedziny czasu przy użyciu Dyskretnej Transformacji Fouriera. Zastosowano następujące charakterystyki widmowe: środek ciężkości widma gęstości mocy (*Audio Spectrum Centroid* - ASC), wariancję środka ciężkości widma gęstości mocy (*Variance of Audio Spectrum Centroid* - varASC), odchylenie średniokwadratowe widma gęstości mocy (*Audio Spectrum Spread* - ASSp), wariancję odchylenia średniokwadratowego widma gęstości mocy (*Variance of Audio Spectrum Spread* - varASSp), skośność odchylenia średniokwadratowego widma gęstości mocy (*Audio Spectrum Skewness* ASSk), wariancję odchylenia średniokwadratowego widma gęstości mocy (*Variance of Audio Spectrum Skewness* - varASSk). [9][10]. Jak wspomniano wcześniej, w rozpoznawaniu sygnału mowy stosuje się najczęściej współczynniki mel-cepstralne [11][12]. W niniejszym artykule wykorzystano 13 pierwszych współczynników MFCC (współczynniki powyżej 13 reprezentują szybkie zmiany współczynników

banku filtrów i nie są przydatne w ASR [13]), ich wariancje, pochodne pierwszego i drugiego rzędu [13]. Za pomocą wzoru (3.1) można obliczyć pochodne pierwszego i drugiego rzędu:

$$d_t = \frac{\sum_{n=1}^{2} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{2} n^2} \quad (3.1)$$

gdzie: *c* - *n*ty współczynnik cepstralny i współczynnik dynamiczny pierwszego rzędu,
t - czas.

4. KLASYFIKACJA

4.1. Sieć neuronowa

W ostatnich latach w rozpoznawaniu mowy i obrazów obserwuje się popularność podejścia "black-box" z wykorzystaniem głębokiego uczenia. Podejście tego typu stało się podstawą do rozwoju wielu komercyjnych systemów detekcji głosu i twarzy. Zwiększenie mocy obliczeniowej i wykorzystanie kart graficznych do przyspieszenia obliczeń równoległych przyczyniło się do stworzenia bardziej złożonych topologii sieci neuronowych do przetwarzania ogromnych ilości danych. Tymczasem eksperymenty wykorzystujące dane FMC zapewniają niezbyt bogate źródło danych. Dlatego w przypadku danych wykorzystanych w prezentowanych badaniach zdecydowano się na zastosowanie prostej architektury sieci neuronowych typu *feed-forward*. Zastosowano sieć neuronową wykorzystującą jedną ukrytą warstwę składającą się z 30 neuronów, korzystającą z funkcji aktywacji Relu (ang. *Rectified Linear Unit*). Wielkość ostatniej warstwy jest określona przez rozmiar klas będących przedmiotem klasyfikacji. Sieć została zaimplementowana przy użyciu biblioteki Keras w języku programowania Python [14].

Tablica 1. Wyniki klasyfikacji z użyciem sieci neuronowej i prostej parametryzacji wysokość-szerokość

Mówca	Sieć Neuronowa (WS)
1	39,6%
2	33,4%
3	38,5%
4	20,1%
5	29,4%
6	37,2%

Tablica 2. Zestawienie wyników klasyfikacji danych FMC dla poszczególnych klasyfikatorów

Mówca	Sieć Neuronowa	SVM (POLY)	SVM (RBF)
1	63,1%	60%	71,4%
2	47,1%	40%	42,9%
3	61,4%	62,9%	60%
4	35,1%	28,6%	34,3%
5	52%	41,1%	32,4%
6	44,5%	88,6%	54,3%

Do treningu wykorzystano algorytm spadku gradientu (ang Stochastic Gradient Descent) o współczynniku uczenia 0,01. Dane zostały wstępnie przetworzone przy użyciu standardowej procedury wyrównywania wartości średniej zbioru danych do zera (ang. *zero-centering)* i normalizacji do przedziału [-1,1]. Do ewaluacji użyto 10-krotnej walidacji krzyżowej, gdzie dla każdej iteracji wybierano model z najlepszym wynikiem. Następnie model ten był wykorzystywany do testów. Trening trwał 1500 epok. W tablicy numer 2 przedstawiono średnie wartości wszystkich iteracji.

4.2 Maszyna wektorów nośnych

W przypadku użycia maszyny wektorów nośnych (ang. *Support Vector Machine*) nie znaleziono opisanego w literaturze zastosowania do klasyfikacji nagrań FMC. Zatem podejście do tego tematu jest nowatorskie. Autorzy musieli sami przetestować empirycznie różne modele i ustawienia tego klasyfikatora zanim uzyskano przedstawione w tej pracy rezultaty. Wykorzystano wcześniejsze dokonania autorów w dziedzinie klasyfikacji głosek alofonicznych przy użyciu sieci neuronowych podczas tworzenia tego klasyfikatora [6][15]. Zastosowany w tej pracy klasyfikator wykorzystujący maszynę wektorów nośnych został zaimplementowany przy użyciu biblioteki scikit-learn w języku programowania Python [16]. Zastosowany moduł przetwarzania danych był identyczny z użytym dla sieci neuronowych.

Wykorzystano dwa rodzaje jądra: jądro wielomianowe oraz jądro RBF (*Radial Basis Function*). Oprócz tego zdecydowano się na zastosowanie wag dla użytych klas ze względu na ich różnorodną liczebność w celu ich zrównoważenia. Posłużono się również automatycznie dobieranym współczynnikiem gamma, natomiast kształt funkcji decyzyjnej przyjęto jako jeden vs reszta. W przypadku użycia jądra wielomianowego skorzystano z 3. stopnia tego jądra. Największym utrudnieniem związanym z tym klasyfikatorem było jego dostrojenie. Ze względu na specyficzne dane, było to zadanie nietrywialne. Dodatkowo ilość danych oraz ich zaszumienie powodowały trudności w poprawnej klasyfikacji.

5. PODEJŚCIE BIMODALNE

W tej części badań oceniono czy dane FMC mogą poprawić rozpoznawanie samogłosek w podejściu wielomodalnym. Do parametryzacji dźwięku użyto standardowych funkcji MFCC, które zostały opisane wcześniej. Algorytm decyzyjny wykorzystujący tylko sygnał audio wykorzystuje współczynniki MFCC w wektorze cech.

Tablica 3. Wyniki klasyfikacji z użyciem dwóch modalności i sieci neuronowej

Mówca	SN (FMC)	SN(AUD)	SN (FUZ)
1	63,1%	76%	86,3%
2	47,1%	80,3%	86,3%
3	61,4%	86%	92,6%
4	35,1%	84,9%	85,4%
5	52,1%	75,6%	78,5%
6	44,6%	82%	85,4%
SUMA:	49,6%	81,2%	85,9%

Tablica 4. Wyniki klasyfikacji z użyciem dwóch modalności i maszyny wektorów nośnych

Mówca	SVM1 FMC	SVM1 AUD	SVM1 FUZ	SVM2 FMC	SVM2 AUD	SVM2 FUZ
1	60%	25,7%	80%	71,4%	60%	77,1%
2	40%	28,6%	74,3%	42,9%	71,4%	74,3%
3	62,9%	25,7%	80%	60%	77,1%	82,9%
4	28,6%	54,3%	71,4%	34,3%	82,8%	74,3%
5	41,2%	64,7%	73,5%	32,4%	70,6%	73,5%
6	45,7%	51,4%	88,6%	54,3%	68,6%	91,4%
SUMA	43,5%	40%	77,2%	48,6%	71%	75,7%

Testowano oba klasyfikatory w 3 ustawieniach: tylko przy użyciu wektora cech audio, tylko wektora FMC oraz z wykorzystaniem obu modalności. Wyniki klasyfikacji przedstawiono w tablicy 3 dla sieci neuronowej oraz dla tablicy 4 dla maszyny wektorów nośnych.

6. WNIOSKI

Na podstawie przeprowadzonych eksperymentów zaobserwowano, że klasyfikator wykorzystujący sieć neuronową jest bardziej skuteczny, niż maszyna wektorów nośnych. SVM przy takich samych ustawieniach algorytmów nie jest w stanie dobrze separować dane audio i FMC, z kolei jednak jest w stanie klasyfikować ich fuzję danych z nieznacznie słabszą skutecznością niż sieć neuronowa. Niestety brak większej ilości danych uniemożliwia korzystanie z podejścia typu black-box i wykorzystania głębokiego uczenia, które mogłoby przynieść znacznie lepsze rezultaty. Jak wspomniano wcześniej, brak danych wynika z faktu, iż zarówno czas nagrań, jak i proces etykietyzacji nagrań są bardzo czasochłonne.

Należy zauważyć, że podejście bimodalne zwiększa skuteczność rozpoznawania alofonów o kilka procent. Jest to bardzo istotne w przypadku posiadania niewielkiej bazy danych. Dzięki takiemu podejściu można łatwo zwiększyć dokładność rozpoznawania głosek alofonicznych.

Kolejnym wnioskiem wynikającym z przeprowadzonych badań jest istotna rola oświetlenia w nagraniach FMC. Odpowiednie doświetlenie mówców pozwoliłoby na uniknięcie zaszumienia danych pochodzących z tej modalności. Szczególnie widoczne jest to dla przypadku mówcy nr 4. Gorszy wynik klasyfikacji najprawdopodobniej jest spowodowany zmianami warunków oświetleniowych w pomieszczeniu.

W przyszłości autorzy pracy mają zamiar wykorzystać inne rodzaje klasyfikatorów. Obiecującym kierunkiem ze względu na parametry audio powinny być Ukryte Modele Markowa (HMM – *Hidden Markov Model*). Wyzwaniem pozostaje dostosowanie parametrów FMC do wymagań tego klasyfikatora oraz wykorzystanie podejścia bimodalnego.

7.PODZIĘKOWANIA

Badania finansowane przez Narodowe Centum Nauki Dec. Nr 2015/17/B/ST6/01874

8. BIBLIOGRAFIA

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, p. 436, May 2015.

[2] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Appl. Intell.*, vol. 42, no. 4, pp. 722–737, 2015.

[3] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of Context-Dependent Deep Neural Networks for Automatic Speech Recognition," *IEEE Spok. Lang. Technol. Work.*, pp. 366–369, 2012.

[4] J. S. Chung and A. Zisserman, "Lip Reading in the Wild."

[5] D. Jachimski, A. Czyzewski, and T. Ciszewski, "A comparative study of English viseme recognition methods and algorithms," *Multimed. Tools Appl.*, 2017.

[6] S. Cygert, G. Szwoch, S. Zaporowski, and A. Czyzewski, "Vocalic Segments Classification Assisted by Mouth Motion Capture," in *2018 11th International Conference on Human System Interaction (HSI)*, 2018, pp. 318–324.

[7] K. B. Korvel G., "Examining Feature Vector for Phoneme Recognition," in *Proceeding of IEEE International Symposium on Signal Processing and Information Technology*, 2017.

[8] A. C. B Kostek, M Piotrowska, T Ciszewski, "No Comparative Study of Self-Organizing Maps vs Subjective Evaluation of Quality of Allophone Pronunciation for Non-native English Speakers," in *Audio Engineering Society Convention 143*, 2017.

[9] B. Kostek *et al.*, "Report of the ISMIS 2011 Contest: Music Information Retrieval," in *Foundations of Intelligent Systems*, 2011, pp. 715–724.

[10] S. T. Hyoung-Gook K., Moreau N., *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. Wiley & Sons, 2005.

[11] D. Eringis and G. Tamulevičius, "Modified Filterbank Analysis Features for Speech Recognition," vol. 3, no. 1, pp. 29–42, 2015.

[12] F. Zheng, G. Zhang, and Z. Song, "Comparison of Different Implementations of MFCC," vol. 16, no. 6, pp. 1–7, 2001.

[13] G. Korvel, O. Kurasova, and B. Kostek, "Comparative Analysis of Spectral and Cepstral Feature Extraction Techniques for Phoneme Modelling," in *Multimedia and Network Information Systems*, 2019, pp. 480–489.

[14] F. Chollet, "Keras." 2015.

[15] S. Zaporowski and A. Czyżewski, "Selection of Features for Multimodal Vocalic Segments Classification," in *Multimedia and Network Information Systems*, 2019, pp. 490–500.

[16] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2012.

REGCORDING, PARAMETERIZATION AND CLASSIFICATION OF ALLOPHONES EMPLOYING BIMODAL APPROACH

The paper concerns the recording and parameterization of allophones in English using two modalities. In the research, the English speakers' statements were recorded. Those speakers's language proficiency corresponds to the level of the native speaker. In the next stage, allophones from audio recordings and corresponding visual signals were isolated. In the process of creating feature vectors, separate parameterization systems were used for each modality. For the audio signal parameterization, typical descriptors used in the area of speech and music recognition were chosen. In the case of the motion capture system own solutions were proposed. For the purpose of allophones classification, neural networks and the suport vector machine were used in both approaches. It has been found that the recognition efficiency increases with the use of more than one modality.

Keywords: Neural networks, classification process, facial motion capture.