XIX Seminarium

ZASTOSOWANIE KOMPUTERÓW W NAUCE I TECHNICE 2019
Oddział Gdański PTETiS

doi: 10.32016/1.68.13

ANALIZA PARAMETRÓW SYGNAŁU MOWY W KONTEKŚCIE ICH PRZYDATNOŚCI W AUTOMATYCZNEJ OCENIE JAKOŚCI EKSPRESJI ŚPIEWU

Szymon ZAPOROWSKI[1], **Bożena KOSTEK**[2]

1. Katedra Systemów Multimedialnych, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska
   tel.: 58 348 63 32 e-mail: smck@multimed.org
2. Laboratorium Akustyki Fonicznej, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska
   tel.: 58 347 16 36 e-mail: bokostek@audioakustyka.org

Streszczenie: Praca dotyczy podejścia do parametryzacji w przypadku klasyfikacji emocji w śpiewie oraz porównania z klasyfikacją emocji w mowie. Do tego celu wykorzystano bazę mowy i śpiewu nacechowanego emocjonalnie RAVDESS (*Ryerson Audio-Visual Database of Emotional Speech and Song*), zawierającą nagrania profesjonalnych aktorów prezentujących sześć różnych emocji. Następnie obliczono współczynniki mel-cepstralne (MFCC) oraz wybrane deskryptory niskopoziomowe MPEG 7. W celu selekcji cech, posiadających najlepsze wyniki rankingowe, wykorzystano las drzew. Następnie dokonano klasyfikacji emocji z za pomocą maszyny wektorów nośnych (SVM, *Support Vector Machine*). Stwierdzono, że parametryzacja skuteczna dla mowy nie jest skuteczna dla śpiewu. Wyznaczono podstawowe parametry, które zgodnie z otrzymanymi wynikami pozwalają na znaczną redukcję wymiarowości wektorów cech, jednocześnie podnosząc skuteczność klasyfikacji.

Słowa kluczowe: niskopoziomowe deskryptory sygnału, analiza śpiewu, ekstrakcja parametrów, klasyfikacja emocji w śpiewie.

1. WSTĘP

Parametryzacja sygnału mowy jest jednym z najlepiej rozwiniętych obszarów przetwarzania sygnałów. Stanowi podstawę szeroko rozumianej komunikacji człowiek-komputer. Jest zwykle pierwszym i często najważniejszym blokiem automatycznego rozpoznawania mowy w połączeniu z algorytmami uczenia maszynowego. Dopiero w ostatnich kilku latach głębokie uczenie wymusiło inne podejście do sygnału mowy, w którym nie poszukuje się najbardziej istotnych parametrów, ale wykorzystuje się najczęściej nieprzetworzony sygnał w postaci obrazów 2D (tj. spektrogramy, cepstrogramy, mel-cepstrogramy, itd.) [1], [2]. Parametry sygnału mowy są również przydatne w przetwarzaniu śpiewu, jednak automatyczna ocena jakości śpiewu w kontekście jego wytwarzania (np. ocena intonacji i barwy głosy śpiewaczego) jest stosunkowo słabo badanym zagadnieniem [3]. Należy przy tym pamiętać, że śpiew – podobnie jak mowa – jest również narzędziem wyrażania uczuć i emocji.

Obszar detekcji emocji w mowie jest dosyć dobrze zbadany, w przeciwieństwie do detekcji emocji w śpiewie. W artykule przedstawiono zagadnienia związane z poszukiwaniem parametrów sygnału mowy, które mogą sprawdzić się w kontekście automatycznej oceny jakości ekspresji w śpiewie. W tym celu wykorzystano bazę nagrań mowy emocjonalnej i śpiewu nacechowanego emocjonalnie, a następnie dokonano parametryzacji tych sygnałów. W kolejnym kroku wyznaczone parametry zostały poddane ocenie z wykorzystaniem algorytmu istotności cech za pomocą lasu drzew. Następnie dokonano klasyfikacji przy użyciu maszyny wektorów nośnych (SVM), bazując na przygotowanych wektorach cech. W końcowej części artykułu przedstawiono wnioski dotyczące rozwoju zaproponowanej metodyki w celu wykorzystania metod uczenia maszynowego do automatycznej oceny jakości ekspresji śpiewu.

2. PRZEGLĄD LITERATURY

2.1. Detekcja emocji

Detekcja emocji w mowie jest obecnie bardzo obecna w literaturze, zwłaszcza kiedy pojawiła się możliwość wykorzystania uczenia głębokiego do tego celu. Większość artykułów opisuje podejścia wykorzystujące sieci neuronowe jako klasyfikatory (sieci splotowe, rekurencyjne, autoenkodery), przedstawiając na wejście algorytmów przetworzone spektrogramy [1], [2], [4]. Stosowanie klasycznych deskryptorów sygnału mowy (np. MFCC) jest obecnie rzadziej spotykane ze względu na niską dokładność rozpoznawania emocji (powyżej 50%) [5], [6]. W przypadku stosowania obrazów jako parametrów skuteczność klasyfikacji może sięgać powyżej 80% [1]. Taką skuteczność można osiągnąć również dla niektórych emocji z użyciem maszyny wektorów nośnych [2].

2.2. Systemy ewaluacji śpiewu

Istnieją systemy pozwalające na automatyczną ocenę jakości śpiewu. Takie systemy skupiają się na ocenie jakości śpiewania pojedynczych głosek lub konkretnej techniki śpiewu [7], [8]. W przypadku tego typu systemów dokładność klasyfikacji może wynosić nawet 80%. Innym stosowanym podejściem jest wykorzystanie częstotliwości podstawowej i badanie czy osoba śpiewająca powtarzająca zadany dźwięk za systemem jest w stanie go poprawnie zaśpiewać [9].

3. BAZA DANYCH I WYBÓR PARAMETRÓW

3.1. Dane

Do przeprowadzenia eksperymentów przedstawionych w niniejszej pracy wykorzystano bazę nagrań RAVDESS.

Baza zawiera nagrania 24 profesjonalnych aktorów (12 kobiet, 12 mężczyzn), śpiewających i wypowiadających dwie dopasowane do siebie wypowiedzi w języku angielskim z neutralnym akcentem północnoamerykańskim. Mowa obejmuje wyrażenia spokoju, radości, smutku, gniewu, strachu, zaskoczenia i zniesmaczenia, a śpiew zawiera emocje spokoju, radości, smutku, gniewu i lęku. Każde wyrażenie jest śpiewane i wymawiane na dwóch poziomach intensywności emocjonalnej (normalna, wzmocniona). Dodatkowo nagrane zostało neutralne emocjonalnie wyrażenie dla każdej z fraz. Przykładowe zestawienie aktorów prezentujących zestaw emocji dostępny w bazie przedstawiono na rysunku 1.

Rys. 1. Przykładowe ekspresje emocji z bazy RAVDESS [10]

Wszystkie nagrania emocji są dostępne w trzech modalnościach: sygnał foniczny (rozdzielczość 16-bitów, częstotliwość próbkowania 48 kHz, pliki w formacie wave), audio-wideo (rozdzielczość 720 p, kodowanie video H.264, kodowanie audio AAC, częstotliwość próbkowania 48 kHz, format pliku mp4) oraz sygnał wideo. Baza zawiera 7356 plików (24,8 GB), z czego 1440 nagrań samej mowy oraz 1012 nagrań śpiewu. Baza jest dostępna na licencji Creative Commons. W pracy wykorzystane zostały tylko pliki audio.

3.2. Parametryzacja

Do parametryzacji danych z bazy RAVDESS wykorzystano dwa podejścia. Wektor parametrów w pierwszym scenariuszu składa się z 40 kolejnych znormalizowanych współczynników mel-cepstralnych (ang. *Mel Frequency Cepstral Coefficients*, MFCC). W przypadku drugiego podejścia, w którym wykorzystano deskryptory MPEG 7 oraz parametry dostępne w bibliotece Librosa [11], wektor cech zawiera parametry zarówno z dziedziny czasu i częstotliwości. Parametry w dziedzinie czasu obejmują: liczbę przejść przez zero (*Zero Crossing*, ZC), energię sygnału (*Root Mean Square Energy*, RMS).

Zastosowano następujące deskryptory widmowe: środek ciężkości widma gęstości mocy (*Audio Spectrum Centroid*, ASC) oraz płaskość widma gęstości mocy (*Audio Spectrum Flatness*, ASF). Oprócz tego wykorzystano wbudowany w bibliotekę Librosa parametr opadania widma (*spectral roll-off*) [11]. Opisywany zestaw parametrów jest obliczany zgodnie z wewnętrznymi nastawami biblioteki Librosa.

4. EKSPERYMENTY

4.1. Ranking istotności

W celu redukcji liczby wykorzystywanych w klasyfikacji parametrów i jednoczesnego zwiększenia dokładności klasyfikacji użyto algorytmu selekcji istotności cech (ang. *Feature Importance*). Algorytm z powodzeniem został wykorzystany przez autorów we wcześniejszych publikacjach związanych z klasyfikacją mowy [12]. Algorytm istotności wektora cech oparty jest na innym algorytmie zwanym ekstremalnie losowymi drzewami (ang. *Extremely Randomized Trees*, ERT) [13]. Koncepcja wywodzi się z losowego lasu (ang. *Random Forest*, RT), który zapewnia kombinację predyktorów drzew, tak że każde drzewo zależy od wartości losowego wektora próbkowanego niezależnie i charakteryzuje się tym samym rozkładem dla wszystkich drzew w lesie. Błąd związany z uogólnieniem dla lasów zbliża się do limitu wraz ze wzrostem liczby drzew w lesie. Błąd uogólnienia ERTs zależy od korelacji między drzewami w lesie i od siły poszczególnych drzew w całym zbiorze [14], [15].

W przeprowadzonych eksperymentach wykorzystano implementację algorytmu ERT zawartą w bibliotece scikit-learn w języku Python [16]. Nastawy algorytmu ERT były następujące: *n_estimators*=’240’, *criterion*=’entropy’, *min_samples_split*=2,*min_samples_leaf*=1, *min_weight_ fraction _leaf*=0.1, *max_features*=’auto’, *min_impurity_ decrease* =0.01, *min_impurity_split*=None, *bootstrap*=True, *random_ state*=True, *warm_start*=True, *class_weight* =balanced.

4.2. Klasyfikacja z użyciem maszyny wektorów nośnych

Do klasyfikacji wykorzystano algorytm maszyny wektorów nośnych zaimplementowaną przy użyciu pakietu scikit-learn w języku Python. Nastawy klasyfikatora dobierano eksperymentalnie, ostatecznie najwyższe wyniki dokładności w klasyfikacji dla wszystkich przebadanych rodzajów emocji uzyskano przy zastosowaniu jądra wielomianowego stopnia 3. wraz z parametrem C = 0.1 i równoważeniem wag poszczególnych klas.

5. WYNIKI

Poniżej przedstawiono wyniki rankingowe istotności poszczególnych parametrów oraz wyniki klasyfikacji emocji. Na rys. 2 przedstawiono ranking istotności mowy i śpiewu dla wszystkich emocji z wykorzystaniem współczynników MFCC. Rys. 3 przedstawia istotność współczynników MFCC w zależności od emocji. W tab. 1-5 przedstawiono dokładność klasyfikacji dla poszczególnych emocji z użyciem maszyny wektorów nośnych oraz liczebność parametrów. W tab. 6 przedstawiono wyniki klasyfikacji dla drugiego scenariusza parametryzacji.

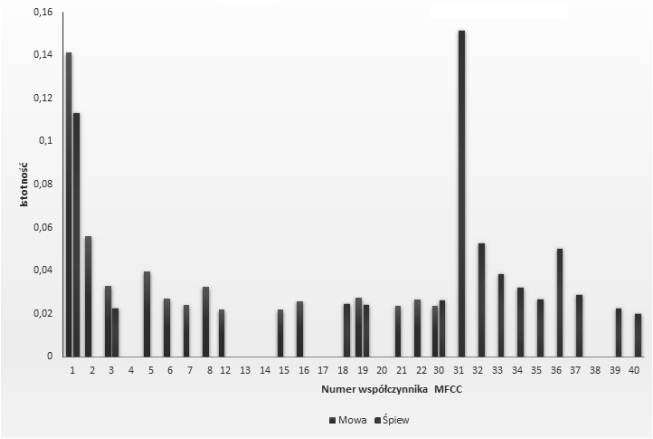
Miara dokładności (wzór 5.1) rozumiana jest poprzez różnicę jedności i stosunku liczby błędnych detekcji w zbiorze do liczby wszystkich przykładów w zbiorze. Błąd średniokwadratowy stanowi wartość oczekiwaną kwadratu błędów. Błędem określana jest różnica pomiędzy wartością uzyskaną za pomocą algorytmu a wartością rzeczywistą. Błąd został przedstawiony we wzorze (5.2).

$$A_{cc} = 1 - \frac{e_{err}}{e} \times 100\% \tag{5.1}$$

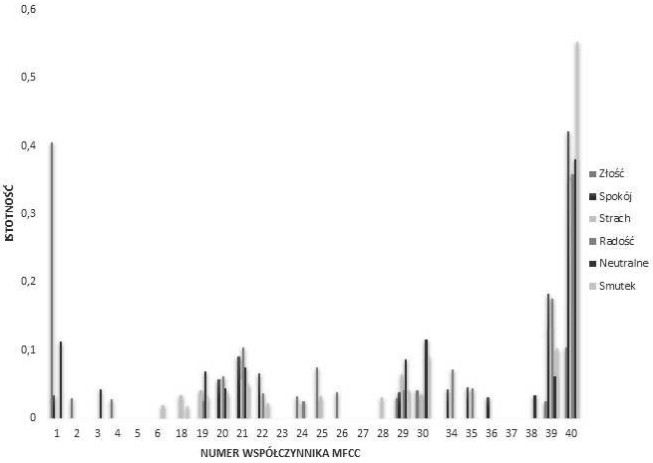
,gdzie e_{err} oznacza liczbę błędnych detekcji, *e* oznacza liczbę wszystkich detekcji

$$e = \frac{1}{n}\sum_{i=1}^{n}(y_i - d_i)^2 \tag{5.2}$$

, gdzie n oznacza kolejne iteracje, y_i oznacza wartość oczekiwaną, a d_i oznacza wartość uzyskaną.



Rys. 2. Wyniki porównania mowy i śpiewu dla unormowanego rankingu istotności współczynników MFCC



Rys. 3. Istotność współczynnika MFCC w zależności od emocji.

Tablica 1. Wyniki klasyfikacji dla emocji: smutek

L. współczynników	Dokładność [%]	Bł. średniokwadr.
40	50,5	0,495
20	73,91	0,2609
15	75,92	0,2408
10	80,93	0,1906
5	87,63	0,124
4	88,63	0,114
2	83,61	0,1639

Tablica 2. Wyniki klasyfikacji mowy i śpiewu dla emocji: złość

L. współczynników	Dokładność [%]	Bł. średniokwadr.
40	67,7	0,323
20	66,81	0,3319
15	68,58	0,3141
10	68,59	0,3142
5	69,47	0,3053
4	70,35	0,2964
2	68,58	0,3142

Tablica 3. Wyniki klasyfikacji mowy i śpiewu dla emocji strach

L.współczynników	Dokładność [%]	Bł. średniokwadr.
40	50,66	0,493
20	50,66	0,4933
15	51,33	0,4867
10	82	0,18
5	90	0,1
4	90,67	0,093
2	69	0,31

Tablica 4. Wyniki klasyfikacji mowy i śpiewu dla emocji spokój

L. współczynników	Dokładność [%]	Bł. średniokwadr.
40	52,67	0,473
20	81,67	0,183
15	84,3	0,157
10	91,7	0,083
5	97	0,03
4	98,3	0,167
2	92,67	0,073

Tablica 5. Wyniki klasyfikacji dla emocji neutralnej

L. współczynników	Dokładność [%]	Bł. średniokwadr.
40	52,7	0,473
20	53,38	0,4662
15	54,05	0,4595
10	70,95	0,29
5	89,19	0,108
4	91,21	0,0878
2	97,973	0,02

Tablica 6. Wyniki klasyfikacji mowy i śpiewu z wykorzystaniem parametrów z biblioteki Librosa

Emocja [%]	ASC	ASF	Roll-off	ZC	RMS
Neutralna	98,52	48,26	34,93	34,53	68,23
Radość	97,87	52,13	38,66	31,81	67,24
Smutek	95,69	47,42	37,84	30,15	62,51
Złość	70,47	27,53	30,92	18,32	47,83
Smutek	93,36	53,77	33,36	24,36	53,27
Strach	96,55	43,29	32,67	21,84	56,68
Wszystkie	79,39	41,23	35,73	27,27	62,26

6. WNIOSKI

Na podstawie przedstawionych wyników można wyodrębnić grupę współczynników MFCC, które są najbardziej istotne w procesie klasyfikacji mowy i śpiewu w obrębie danej emocji. Rozróżnianie emocji w mowie i śpiewie z wykorzystaniem tych współczynników charakteryzuje się wysoką dokładnością dla większości emocji (powyżej 88%). W większości przypadków zredukowany do dwóch cech wektor danych składał się z 29 i 40 współczynnika MFCC. Dla emocji złości były to współczynniki 1 i 39. Spośród przebadanych współczynników z drugiego wariantu wektora cech najwyższy wynik uzyskano z wykorzystaniem środka ciężkości widma gęstości mocy (ASC). Zastanawiająca jest niska skuteczność parametru przejść przez zero (ZC) oraz energii RMS. Na podstawie przeprowadzonych eksperymentów można zaobserwować, iż współczynniki MFCC osiągają znacznie lepsze wyniki klasyfikacji. Wydają się one być naturalnym kierunkiem w dalszych pracach nad systemem oceny jakości ekspresji w śpiewie. Zauważono również spadek dokładności klasyfikacji dla emocji złości we wszystkich wykorzystanych wektorach parametrów. Jest

to interesujące zjawisko, które powinno być zbadane w oparciu o inną bazę nagrań. Taka różnica może wynikać ze względu na znaczną zmianę głośności wypowiedzi oraz możliwe zmiany częstotliwości formantowych w przypadku tej emocji. Sam sposób artykułowania powiązany z emocją również może wpływać na dokładność klasyfikacji. Dokładność klasyfikacji pozostałych emocji jest zbliżona do siebie.

W przyszłości autorzy pracy mają zamiar skupić się na stworzeniu parametryzacji w oparciu o wszystkie deskryptory niskopoziomowe MPEG-7 i sprawdzenie ich skuteczności w klasyfikacji emocji zarówno w mowie, jak i śpiewie. Kolejnym krokiem będzie również testowanie parametryzacji na zbiorach zawierających śpiew operowy.

W literaturze nie ma odniesień do systemów oceniających jakość ekspresji w śpiewie. Podstawą takiego systemu mogłaby być detekcja emocji w śpiewie, rozbudowana o system rankingowy, wykorzystująca podejście opisane w artykule, chociaż naturalne wydaje się rozszerzenie badań w kierunku zastosowania uczenia głębokiego i reprezentacji 2D sygnałów.

7. BIBLIOGRAFIA

1. D. Bertero and P. Fung: A first look into a Convolutional Neural Network for speech emotion detection, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, 5115–5119.
2. L. Kerkeni, Y. Serrestou, K. Raoof, C. Cléder, M. Mahjoub, and M. Mbarki: Automatic Speech Emotion Recognition Using Machine Learning, 2019, p. https://www.intechopen.com/online-first/automatic.
3. K. R. Scherer, J. Sundberg, L. Tamarit, and G. L. Salomão: Comparing the acoustic expression of emotion in the speaking and the singing voice, *Comput. Speech Lang.*, vol. 29, no. 1, 218–235, 2015.
4. N. Cibau, E. Albornoz, and H. Rufiner, *Speech emotion recognition using a deep autoencoder*. 2013.
5. M. C. Sezgin, B. Gunsel, and G. K. Kurt: Perceptual audio features for emotion detection, *EURASIP J. Audio, Speech, Music Process.*, vol. 2012, no. 1, p. 16, 2012.
6. S. S. Poorna, C. Y. Jeevitha, S. J. Nair, S. Santhosh, and G. J. Nair: Emotion recognition using multi-parameter speech feature classification, in *2015 International Conference on Computers, Communications, and Systems (ICCCS)*, 2015, 217–222.
7. P. Zwan: Expert system for automatic classification and quality assessment of singing voices, *Audio Eng. Soc. - 121st Conv. Pap. 2006*, vol. 1, 446–454, Jan. 2006.
8. N. Amir, O. Michaeli, and O. Amir: Acoustic and perceptual assessment of vibrato quality of singing students, *BIOMED SIGNAL Process Control*, vol. 1, 144–150, Apr. 2006.
9. E. Półrolniczak and M. Łazoryszczak: Quality assessment of intonation of choir singers using F0 and trend lines for singing sequence, *Metod. Inform. Stosow.*, vol. no. 4, 259–268, 2011.
10. S. R. Livingstone and F. A. Russo, *The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north American english*, vol. 13, no. 5. 2018.
11. B. McFee *et al.*: librosa/librosa: 2019.
12. S. Zaporowski and A. Czyżewski: Selection of Features for Multimodal Vocalic Segments Classification BT - Multimedia and Network Information Systems, 2019, 490–500.
13. P. Geurts, D. Ernst, and L. Wehenkel: Extremely randomized trees, *Mach. Learn.*, vol. 63, no. 1, 3–42, 2006.
14. G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts: Understanding variable importances in forests of randomized trees, *Advances in Neural Information Processing Systems 26* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds.) Curran Associates, Inc., 2013, 431–439.
15. V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston: Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling, *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, 1947–1958, Nov. 2003.
16. F. Pedregosa *et al.*: Scikit-learn: Machine Learning in {P}ython, *J. Mach. Learn. Res.*, vol. 12, 2825–2830, 2011.

ANALYSIS OF THE SPEECH SIGNAL PARAMETERS IN THE CONTEXT OF THEIR SUITABILITY IN THE AUTOMATIC QUALITY OF SINGING EXPRESSION ASSESSMENT

This paper concerns the approach to parameterization for the classification of emotions in singing and comparison with the classification of emotions in speech. For this purpose, the RAVDESS database containing emotional speech and song was used. This database contains recordings of professional actors presenting six different emotions. Next, Mel Frequency Cepstral Coefficients and selected Low-Level MPEG 7 descriptors were calculated. Using the algorithm of Feature Selection based on a Forest of Trees, coefficients, and descriptors with the best ranking results were determined. Then, the emotions were classified using the Support Vector Machine. The classification was repeated several times, and the results were averaged. It was found that descriptors used for emotion detection in speech are not as useful for singing. Basic parameters for singing were determined which, according to the obtained results, allow for a significant reduction in the dimensionality of feature vectors while increasing the classification efficiency of emotion detection.

Keywords: Mel Frequency Cepstral Coefficients. MPEG 7 Low-Level Audio Descriptors, singing analysis, Feature Selection.