Post-print of: Zaporowski S., Kostek B. (2020) Ranking Speech Features for Their Usage in Singing Emotion Classification. In: Helic D., Leitner G., Stettinger M., Felfernig A., Raś Z.W. (eds) Foundations of Intelligent Systems. ISMIS 2020. Lecture Notes in Computer Science, vol 12117. Springer, Cham. https://doi.org/10.1007/978-3-030-59491-6_21

Ranking Speech Features for their Usage in Singing Emotion Classification

Szymon Zaporowski[1] [0000-0003-0814-1097] and Bozena Kostek[2][0000-0001-6288-2908]

[1] Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics, Multimedia Systems Department, 80-233 Gdansk, Poland
[2] Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics, Audio Acoustics Laboratory, 80-233 Gdansk, Poland
smck@multimed.org, bokostek@audioacoustics.org

Abstract. This paper aims to retrieve speech descriptors that may be useful for the classification of emotions in singing. For this purpose, Mel Frequency Cepstral Coefficients (MFCC) and selected Low-Level MPEG 7 descriptors were calculated based on the RAVDESS dataset. The database contains recordings of emotional speech and singing of professional actors presenting six different emotions. Employing the algorithm of Feature Selection based on the Forest of Trees method, descriptors with the best ranking results were determined. Then, the emotions were classified using the Support Vector Machine (SVM). The training was performed several times, and the results were averaged. It was found that descriptors used for emotion detection in speech are not as useful for singing. Also, an approach using Convolutional Neural Network (CNN) employing spectrogram representation of audio signals was tested. Several 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 (MFCC), MPEG 7 Low-Level Audio Descriptors, feature selection, singing expression classification

1 Introduction

Speech analysis and processing, parametrization as well as automatic classification are the areas being thoroughly researched and developed for the last few decades as their application is of utmost importance in many domains. To name a few [1]–[4]: telecommunications (VoIP, enhanced IP communication services), automatic speech transcription, automated speech-to-text technologies in video-over IP communications, medical applications such as hearing aids, cochlear implants and speech pathology recognition, language processing for communication services, and more recently human-(intelligent) computer communication based on big data [5, 6]. The last-mentioned application is within the interest of researches as well as commercial usage.

Parametrization is usually the first and often the most crucial block of automatic speech recognition (ASR) in combination with machine learning algorithms. It is only in the last few years that deep learning methodology has forced a different approach to the speech signal processing, in which speech parameters are not retrieved, but the signal in the form of 2D images (i.e., spectrograms, cepstrograms, mel-cepstrograms, chromagrams, *wavenet*-like, etc.) [7]–[9] is fed at the net input. On the other hand, automatic evaluation of singing quality in the context of its production (e.g., evaluation of the intonation and timbre of the singing voice) is a relatively poorly studied issue comparing to the 'pure' speech area [3, 10]. It should, however, be remembered that singing - like speech - is also a tool for expressing feelings and emotions, thus speech descriptors applied to the singing evaluation should be useful. Also, it is interesting whether deep learning-based methodology may be – in a straightforward way - applied to the singing expression evaluation.

The area of emotion detection in speech is quite well studied, in contrast to the detection of emotion in singing. The article presents issues related to the search for speech signal parameters that may work in the context of automatic evaluation of the quality of expression in singing. For this purpose, a dataset containing recordings of emotional speech and emotionally-singing singing was used, followed by the parametrization of these signals. Some speech descriptors were evaluated for their usage in the feature vector (FV) for singing emotion recognition.

In the next step, the determined parameters were evaluated and reduced using the feature significance algorithm using a Forest of Trees method. Then classification was carried out using the Support Vector Machine (SVM) based on the prepared reduced feature vectors. The final part of the article presents conclusions regarding the development of the proposed methodology to use machine learning methods, including a deep learning approach, to assess the quality of singing expression automatically.

2 Related Work

2.1 Emotion Detection in Speech

The detection of emotions in speech is now very much present in the literature, especially when the possibility of using deep learning for this purpose appeared. Most of the studies describe approaches that use artificial neural networks as classifiers (i.e., convolutional networks, recursive networks, autoencoders), presenting processed spectrograms as input [7, 8, 11]. The use of classical speech signal descriptors (e.g., Mel-Frequency-Cepstral-Coefficients, MFCC) is currently less prevalent in speech research due to the lower accuracy of emotion recognition (approx. 60%) [12, 13]. When 2D image spaces are used as parameters, the classification efficiency can reach over 80% [7, 9]. Such efficiency can also be achieved for some chosen emotions using SVM [8].

2.2 Emotion in Singing

There exist systems that allow automatic assessment of singing and singing quality. The focus of such systems is on assessing the quality of singing individual sounds or a specific singing technique [14, 15]. Classification accuracy can be up to 80% for these types of systems. Another approach researched is to use the fundamental frequency as a parameter to test whether the person singing a given sound or repeating it after the system prompt is able to sing it correctly [16].

3 Parameter Selection

3.1 Dataset

The RAVDESS database of recordings was used to conduct the experiments presented in this paper. This dataset is often used in research studies, thus it may be treated as a benchmark in the area of speech emotion recognition. The database contains recordings of 24 professional actors (12 women, 12 men) speaking and singing two matched English statements with a neutral North American accent. Speech includes expressions of calm, joy, sadness, anger, fear, surprise and disgust, and singing contains the emotions of calm, joy, sadness, anger, and fear. Each expression is sung and pronounced at two levels of emotional intensity (normal, enhanced). Additionally, an emotionally neutral expression was recorded for each phrase. An example of actors' images presenting a set of emotions available in the database is shown in Fig. 1.

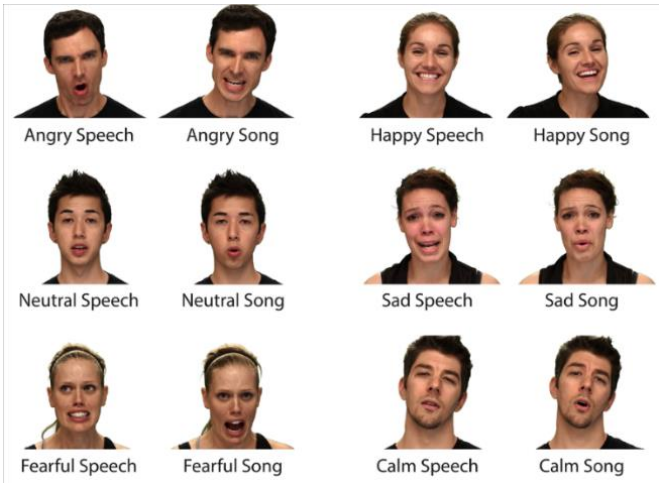

Fig. 1. Example of the RAVDESS emotion expression [17]

All emotion recordings are available in three modalities: audio signal (16-bit resolution, 48 kHz sampling frequency, wave format files), audio-video (720 p resolution, H.264 video coding, AAC audio coding, 48 kHz sampling frequency), mp4 file for-

mat) and video signal. The database contains 7356 files (24.8 GB), of which 1440 recordings are of speech alone, and 1012 recordings are of singing. The database is available under a Creative Commons license. Only audio files were used in this study.

3.2 Parameter Selection

Two approaches were utilized to parameterize data from the RAVDESS database. The FV in the first scenario consists of 40 consecutive normalized MFCCs. In the second approach, which uses MPEG 7 descriptors and parameters available in the Librosa library [18], FV contains parameters in both time- and frequency-domains. Time-domain parameters include zero crossings (Zero-Crossing, ZC), and signal energy (Root Mean Square Energy, RMS). The following spectral descriptors were used [18]: the spectral center of gravity (Audio Spectrum Centroid, ASC) and the spectral flatness measure (Audio Spectrum Flatness, ASF). Besides, the spectral roll-off parameter built into the Librosa library was employed [18]. This set of parameters is calculated according to the internal settings of the Librosa library. All the above-mentioned descriptors are present in the literature [19, 20]; thus their definitions will not be recalled here.

For the approach based on Convolutional Neural Networks (CNN), spectrograms were calculated for each of the utterances and songs from the RAVDESS corpora. The audio is sampled at 48000 Hz. Each audio frame is windowed using the Kaiser window of the length of 2048. Fast Fourier Transform (FFT) windows of the length of 2048 are then applied on the windowed audio samples with the STFT hop-length as 512 points. As a result of the aforementioned transformations, the bandwidth of the audio signal was reduced to 8 kHz. In total, there were more than 24200 samples for six classes. That means there were more than 4050 examples for each class.

4 Experiments

4.1 Significance Ranking

To reduce the number of parameters used in the classification and, at the same time, increase the accuracy of the classification by leaving only significant descriptors, the Feature Importance algorithm was employed. The authors have successfully utilized this algorithm in earlier publications related to speech classification [21]. The Feature Importance algorithm is based on another algorithm called Extremely Randomized Trees (ERT) [22]. The concept is derived from Random Forest (RT), which provides a combination of tree predictors so that each tree depends on the value of a random vector sampled independently and has the same distribution for all trees in the forest. The error related to generalization for forests is approaching the limit as the number of trees in the forest increases. ERTs generalization error depends on the correlation between trees in the forest and the strength of individual trees in the entire set [21, 22].

The conducted experiments used the implementation of the ERT algorithm contained in the scikit-learn library in Python [25]. The ERT algorithm settings were as

follows: 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 SVM-based Classifier

The SVM algorithm, using the scikit-learn package in Python, was employed for the classification. The classifier settings were selected experimentally, ultimately the highest accuracy in the classification for all types of emotions studied was obtained using a degree 3 polynomial kernel with the parameter C = 0.1 and 'balanced' mode of adjusting weights of individual classes. For comparing classes with each other, one vs. all approach was used.

4.3 CNN Classifier

The CNN classifier used for this experiment was created using the Tensorflow library. The architecture used for this experiment is shown in Fig. 2. The architecture was created in an empirical approach, adding individual layers, and then examining their impact on classification results. Inspiration for this architecture was research presented in the literature [24, 25]. The created neural network was trained for 200 epochs using a batch size of 32 and data split 60/40 for training and validation set, respectively. Titan RTX graphic processor was employed for training.

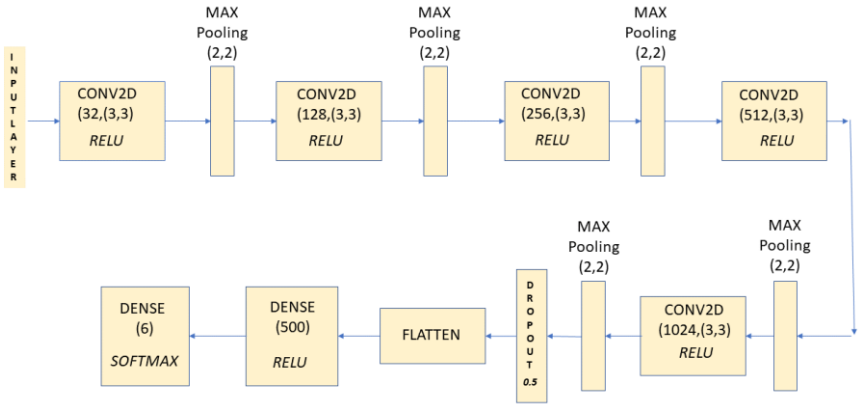

Fig. 2. Architecture of CNN used in the presented experiment

5 Results and discussions

Below ranking results of the significance of individual parameters and results of emotion classification are shown. Fig. 3a) shows the importance of the MFCCs depending

on emotions. According to Fig. 3a) coefficient no. 40 is the most versatile. It is the most important feature for several emotions, e.g., joy, calm, sadness, and neutral state. Also, feature no. 1 seems to be most important for anger and fear emotions. Fig. 3b) presents the ranking of the parameter importance of speech and singing for all emotions using the MFCC coefficients. As can be seen in Fig. 3b) the most essential parameters for speech and singing are different. The common one is the coefficient no. 1, but the rest differs. For speech, more important are coefficients with the lower numbers of order, in the case of singing, features with the number higher than 30 were indicated. Tabs. 1-3 show several MFCC parameters, the accuracy of classification as well as the mean square error (MSE) for individual emotions using SVM. The MFCCs shown are derived from the Feature Importance algorithm, which indicated the most important features respectively to values presented in Tables 1-3. In most cases, the use of four best coefficients provides the best accuracy results. It is worth noticing that reducing FV to only 10 best features in most cases results in a significant increase in the accuracy score. The emotion of anger is an exception here; accuracy values are oscillating all the time around 70%. Tab. 4 contains the classification results for the second parameterization scenario employing MPEG-7 low-level descriptors. Tab 5 presents results for the CNN-based classification approach. The measure of accuracy is understood as the ratio of the number of correct predictions to the total number of input samples [28].

Based on the presented results, it is possible to distinguish a group of MFCC coefficients that are most important in the process of classifying speech and singing within a given emotion. Classifying emotions in speech and singing using these factors is characterized by high accuracy for most emotions (over 88%). In most cases, the feature vector reduced to two descriptors consisted of MFCCs nos. 29 and 40. For anger, these were coefficients nos. 1 and 39. Among the tested coefficients from the second FV variant, the highest result was obtained using spectral centroid (ASC). The low efficiency of the zero-crossing (ZC) parameter and RMS energy is puzzling. Based on the experiments conducted, it can be observed that MFCC coefficients achieve much better classification results. They seem to be a natural direction in further work on the system for assessing the quality of expression in singing. There was also a decrease in the classification accuracy for anger emotions in all the feature vectors used. This is an interesting phenomenon that should be studied based on another database of recordings. Such a difference may result from a significant change in the volume of speech and possible changes in formant frequencies in the case of this emotion. Articulation associated with emotion can also affect the accuracy of classification. The accuracy of the classification of other emotions is similar. Results for the CNN approach are slightly worse than the results for MFCC parameterization. It could be due to the fact that spectrograms bandwidth were limited to only 8 kHz. It is worth noticing that the categorical cross-entropy values indicate that there is room for improvement, however, presented values and architecture were the best from all tested architectures.

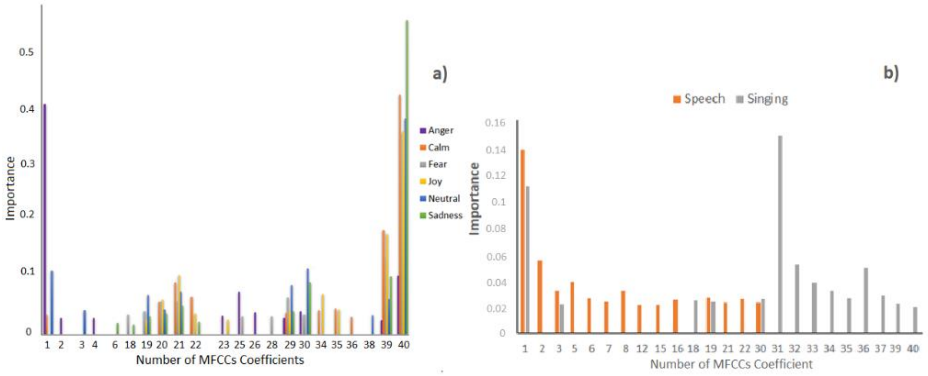

Fig. 3. Classification results for a) The importance of MFCCs in terms of the researched emotion b) for speech and singing for the normalized ranking of MFCCs

Table 1. Classification (speech and singing) results for anger

Quantity of the MFCCs retained	Accuracy [%]	MSE
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

Table 2. Classification (speech and singing) results for fear

Quantity of the MFCCs retained	Accuracy [%]	MSE
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

Table 3. Classification (speech and singing) results for neutral emotion

Quantity of the MFCCs retained	Accuracy [%]	MSE
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**

Table 4. Emotion classification results for speech and singing based on MPEG 7 descriptors

Emotion [%]	ASC	ASF	Roll-off	ZC	RMS
Neutral	**98.52**	48.26	34.93	**34.53**	**68.23**
Joy	97.87	52.13	**38.66**	31.81	67.24
Sadness	95.69	47.42	37.84	30.15	62.51
Anger	**70.47**	27.53	30.92	18.32	47.83
Surprised	93.36	**53.77**	33.36	24.36	53.27
Fear	96.55	43.29	32.67	21.84	56.68
All	79.39	41.23	35.73	27.27	62.26
Average	**90.26**	**44.80**	**34.87**	**26.90**	**59.72**

Table. 5. Emotion classification results for speech and singing based on the CNN approach

Emotion	Accuracy [%]	Categorical Cross-Entropy
Neutral	75.85	0.8693
Joy	**51.33**	7.8441
Sadness	57.83	2.9245
Anger	76.33	0.9483
Surprised	**77.33**	1.9979
Fear	77.00	1.3644
All	**65.96**	1.4873

6 Conclusions

In this paper, an approach to rank speech features based on RAVDESS emotional speech and singing dataset with different approaches to parameterization and classification is presented. Significance of particular MFCC parameters for speech and singing derived from the Feature Importance algorithm is shown. Three different approaches to parameterization using MFCCs, MPEG-7 low-level descriptors, and spectrograms are also demonstrated. The results for each approach are presented and discussed.

In the future, the authors intend to focus on creating parameterization based on all MPEG-7 low-level descriptors and checking their effectiveness in the classification of emotions, both in speech and singing. The next step will also be testing parameterization on sets containing opera singing. The basis of such a system could be the detection of emotions in singing, expanded by a ranking system, using the approach described in the article. However, it seems natural to extend research towards the use of

deep learning and 2D representation of signals such as cochleagrams or CQT (Constant-Q) transform.

References

[1] K. Mukesh and S. . Shimi, "Voice Recognition Based Home Automation System for Paralyzed People," *Int. J. Adv. Res. Electron. Commun. Eng.*, vol. 4, no. 10, pp. 2508–2515, 2015.

[2] J. Markoff, "From Your Mouth to Your Screen, Transcribing Takes the Next Step," 2019. [Online]. Available: https://www.nytimes.com/2019/10/02/technology/automatic-speech-transcription-ai.html. [Accessed: 15-Jan-2020].

[3] A. Munir, S. Kashif Ehsan, S. M. Mohsin Raza, and M. Mudassir, "Face and speech recognition based smart home," *2019 Int. Conf. Eng. Emerg. Technol. ICEET 2019*, pp. 1–5, 2019.

[4] V. Delić *et al.*, "Speech technology progress based on new machine learning paradigm," *Comput. Intell. Neurosci.*, vol. 2019, 2019.

[5] X. Lei, G.-H. Tu, A. X. Liu, K. Ali, C.-Y. Li, and T. Xie, "The Insecurity of Home Digital Voice Assistants -- Amazon Alexa as a Case Study," 2017.

[6] K. Kannan and J. Selvakumar, "Arduino Based Voice Controlled Robot," *Int. Res. J. Eng. Technol.*, vol. 02, no. 01, pp. 49–55, 2015.

[7] 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, pp. 5115–5119.

[8] 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.

[9] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, 2019.

[10] 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, pp. 218–235, 2015.

[11] N. Cibau, E. Albornoz, and H. Rufiner, "Speech emotion recognition using a deep autoencoder," in *Anales de la XV Reunion de Procesamiento de la Informacion y Control*, 2013, pp. 934–939.

[12] 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.

[13] 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, pp. 217–222.

[14] P. Zwan, "Expert system for automatic classification and quality assessment of singing voices," *Audio Eng. Soc. - 121st Conv. Pap. 2006*, vol. 1, pp. 446–454, Jan. 2006.

[15] N. Amir, O. Michaeli, and O. Amir, "Acoustic and perceptual assessment of vibrato

quality of singing students," *Biomed. Signal Process. Control - BIOMED SIGNAL Process Control*, vol. 1, pp. 144–150, Apr. 2006.

[16] E. Półrolniczak and M. Łazoryszczak, "Quality assessment of intonation of choir singers using F0 and trend lines for singing sequence," *Metod. Inform. Stosow.*, vol. nr 4, pp. 259–268, 2011.

[17] 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.

[18] D. Ellis *et al.*, "librosa: Audio and Music Signal Analysis in Python," *Proc. 14th Python Sci. Conf.*, no. Scipy, pp. 18–24, 2018.

[19] G. Muhammad and M. Melhem, "Pathological voice detection and binary classification using MPEG-7 audio features," *Biomed. Signal Process. Control*, vol. 11, no. 1, pp. 1–9, 2014.

[20] N. Dave, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition," *Int. J. Adv. Res. Eng. Technol.*, vol. 1, no. Vi, pp. 1–5, 2013.

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

[22] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006.

[23] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," *Adv. Neural Inf. Process. Syst. 26*, pp. 431–439, 2013.

[24] V. Svetnik, A. Liaw, C. Tong, J. Christopher 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, pp. 1947–1958, 2003.

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

[26] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, "Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions," pp. 1–12, 2019.

[27] X. Liu, Q. Chen, X. Wu, Y. Liu, and Y. Liu, "CNN based music emotion classification," *arXiv Prepr.*, 2017.

[28] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.